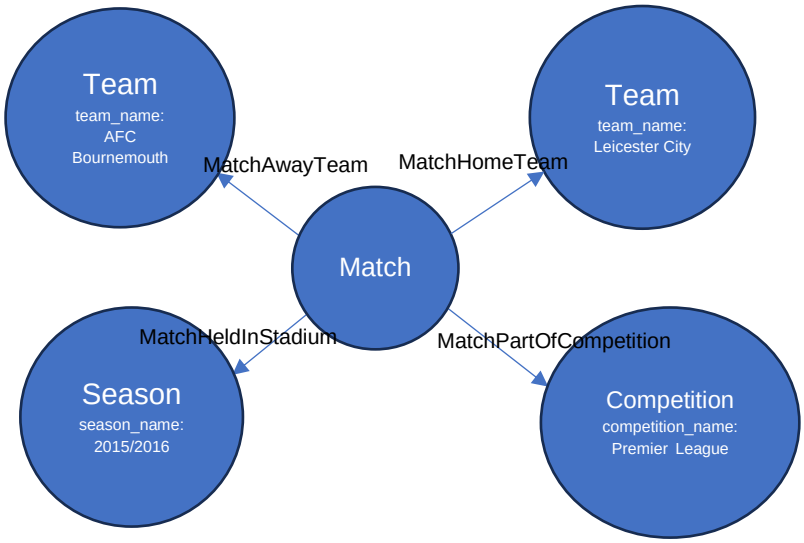
Data Modeling for KickStat Knowledge Graph

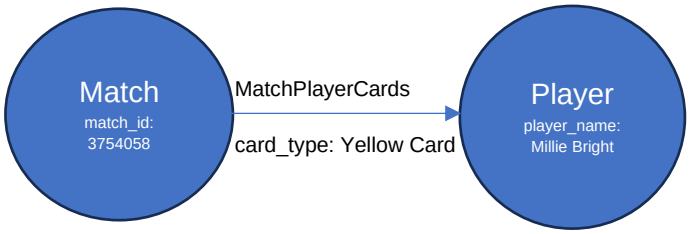
As Statsbomb open data is centered around matches, to efficiently retrieve rare events the following guidelines are followed to build a Knowledge Graph:

Rule 1. "Match" is modeled as a core entity and all simple properties specific to a match like match_id, match_date etc., are modeled as entity properties.

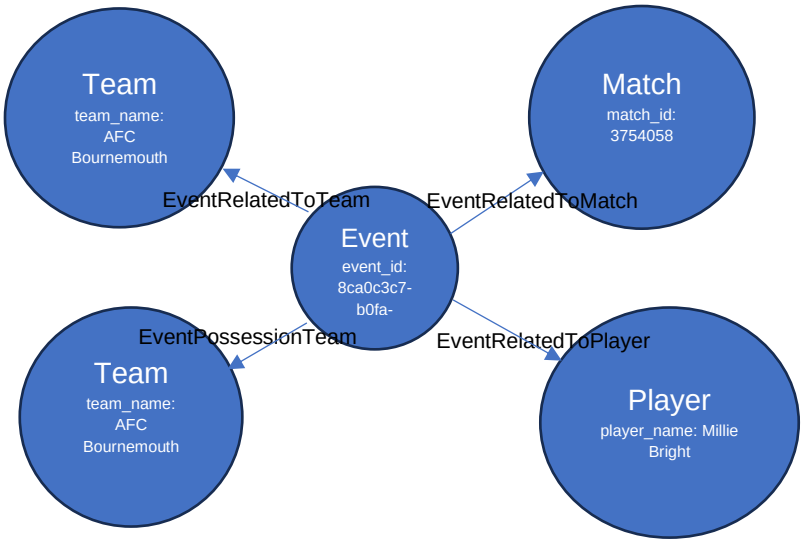
Rule 2. Composite properties of match entity like home_team, competition, season etc., are modeled as separate entities, called supported entities and relations are created to connect these entities with the hosting Match core entity.



Rule 3. Properties which are related to both Match entity and supported entities will be added to the relationship, for example, if a player receives a card in a match, the card type for the player for that match is captured as a relation property and added to the relationship between the player and the match.



Rule 4. Match event data is modeled as Event entities connected to corresponding match and supported entities, for example, possession_team property is connected to Team entity, player property is connected to corresponding Player entity. Also, as event data consists of a large number of properties related to various event types, we had two options to capture the data in the knowledge graph: capturing as properties of event node itself or capturing as subevent entities each specific to a type of event like Bad behavior, Pass etc. The first option comes with the disadvantage that there will be a lot of sparse entries in each event entity node, while the second option comes with the disadvantage that there will be an increased hops in retrieving relevant data. We went with the first option as we are not considering all the event types due to limited scope.



Rule 5. Some Event data like Dribble, Interception etc., due to limited scope and which is not relevant in uncovering rare events is omitted in into the knowledge graph, however, the same data can be handled as event data as mentioned in Rule 4 can be handled the same way.

Rule 6. Match 360 data can also be handled in the same way, however, at this point, as this data is highly independent to other data, and needs significant work to correlate with match data, we do not incorporate 360 data into knowledge graph, and will be dealt in the future.

Detailed list of Entities and Relationships of the Knowledge Graph:

Entities:

Entity	Description	Properties
MatchNode	Entity describing matches data. Note: As Entity is a keyword in Kuzu, we had to name Match entity as MatchNode to avoid conflict.	match_id INT64, match_date DATE, kick_off TIMESTAMP, home_score INT64, away_score INT64, match_status STRING, match_status_360 STRING, match_week INT64, PRIMARY KEY (match_id))
Competition	Entity describing competitions data.	competition_id INT64, country_name STRING, competition_name STRING, PRIMARY KEY (competition_id))
Season	Entity describing seasons data.	season_id INT64,season_name STRING, PRIMARY KEY (season_id))
Stadium	Entity describing stadiums data.	id INT64, name STRING, PRIMARY KEY (id))
Country	Entity describing countries data.	id INT64, name STRING, PRIMARY KEY (id))
Referee	Entity describing referees' data.	id INT64, name STRING, PRIMARY KEY (id))
Managers	Entity describing managers data.	id INT64, name STRING, nickname STRING, dob DATE, PRIMARY KEY (id))
CompetitionStage	Entity describing competition stages data.	id INT64, name STRING, PRIMARY KEY (id))
Team	Entity describing teams' data.	team_id INT64, team_name STRING, team_gender STRING, PRIMARY KEY (team_id))
Player	Entity describing players data.	player_id INT64, player_name STRING, player_nickname STRING, PRIMARY KEY (player_id))
Event	Entity describing events data.	event_id STRING, index INT64, period INT64, timestamp STRING, minute INT64, second INT64, type_id INT64, type_name STRING, possession INT64, play_pattern_id INT64, play_pattern_name STRING, position_id INT64, position_name STRING, location STRING, duration FLOAT, under_pressure BOOLEAN, off_camera BOOLEAN, out BOOLEAN, related_events STRING, bad_behaviour_card_id INT64, bad_behaviour_card_name STRING, PRIMARY KEY (event_id))

Relations

Relation Name	Source and Destination	Description	Relation properties
TeamCountry	FROM Team TO Country		
TeamManagers	FROM Team TO Managers		
MatchAwayTeam	FROM MatchNode TO Team		AwayTeamGroup STRING
MatchHomeTeam	FROM MatchNode TO Team		AwayTeamGroup STRING
MatchPartOfCompetition	FROM MatchNode TO Competition		
MatchHeldInSeason	FROM MatchNode TO Season		
MatchHeldInStadium	FROM MatchNode TO Stadium		
StadiumCountry	FROM Stadium TO Country		
RefereeCountry	FROM Referee TO Country		
ManagerCountry	FROM Managers TO Country		
MatchCompetitionStage	FROM MatchNode TO CompetitionStage		
MatchReferee	FROM MatchNode TO Referee		
MatchPlayers	FROM MatchNode TO Player		jersey_number INT64
PlayerCountry	FROM Player TO Country		
PlayerTeam	FROM Player TO Team		
EventRelatedToMatch	FROM Event TO MatchNode		
EventRelatedToPlayer	FROM Event TO Player		
EventRelatedToTeam	FROM Event TO Team		
EventPossessionTeam	FROM Event TO Team		