Proposal and Design of Hockey Statistics Database

Instructor: Arik Senderovich
Group members: A Mahfouz, Faria Khandaker, Osama Farooq
Winter 2020 Semester

As one of Canada's national sports, ice hockey remains a source of entertainment for millions of households across Canada. While many watch the game for entertainment, others delve much further into the game utilizing data analytics to discover high-quality insights. The emergence of data analytics has caused drastic changes in the hockey world, with some teams deviating from traditional strategies and play-styles, and replacing it with new ones that align with statistics. That said, in order to conduct such analysis, an appropriate platform for storage and retrieval of data is imperative. This report proposes a simple database design for storing data for the National Hockey League to maximize querying efforts and return detailed results for analysis. Specifically, it includes the formulation of an entity-relationship diagram, which is later translated into relations and realized as a MySQL database.

1.1 Application Domain

Domain Description

In the movie Moneyball, General Manager Peter Brand displays the importance of data analytics in sports as he attempts to construct a winning team while faced with a limited budget. The emergence of data analytics in sports is taking teams beyond the traditional metrics for evaluating players by allowing teams to identify and strategically target players of maximum value. The market for this rapidly growing phenomenon is expected to reach almost $4 billion by 2022. (Prewitt, 2019) By understanding player strengths, weaknesses, and trends, general managers and coaches can strategically acquire or release players, with the ultimate goal of constructing a winning roster. Moreover, by analyzing team statistics, coaches can elicit a game plan to exploit the opponent's weaknesses.

As one (1) of the four (4) major sports leagues in North America, the National Hockey Association ("NHL") and its associated teams are not foreign to employing data analytics techniques in order to gain a competitive advantage. That being said, in order to manipulate data for a competitive advantage, an appropriate solution must be developed that is capable of storing large amounts of data while allowing multiple users to access the data simultaneously. For context, the NHL consists of thirty-one (31) teams and each team consists of twenty-three (23) players and one (1) head coach. Accordingly, the NHL records and reports on details regarding 713 players for any given season, which includes the pre-season (exhibition matches), regular season (round-robin format), and postseason (best-of-seven single-elimination series). Furthermore, these records are further augmented with various categories, including but not limited to: rosters, salaries, team statistics, individual match statistics, player statistics, and awards. Due to the sheer size of the available data, and to ensure data integrity, a uniform database management system ("DBMS") is an appropriate solution. This report focuses on creating a database management system for managing National Hockey League ("NHL") statistics. Specifically, this DBMS will store an extensive amount of player and team data, and will allow for multiple users to simultaneously access this data for various purposes including data analysis and fantasy sports. Seeing as sports are becoming increasingly data-driven, by creating a DBMS, stakeholders have access to accurate customizable data and player statistics.

1.2 Sample Queries (10)

1. Which 10 teams have the most points during the 2019-2020 season?
2. How many goals did [team] score at home this season versus away?
3. Who are the goalies with the highest save percentages?
4. Who is the oldest player on each team this season?
5. When are all the Maple Leafs home games this season?
6. What are the names and teams of all defencemen?
7. What are the names of everyone affiliated with [team name] this season?
8. Who has won the Vezina Trophy the last [x] years?
9. Who are the head coaches for all the Eastern Conference teams?
10. What teams has [player name] been associated with?

1.3 Data Requirements and ER Diagram

- All hockey persons can be uniquely identified by their name and birthday. However, for convenience and since birth dates can be unknown, hockey persons will be assigned a unique ID.
- Hockey persons can be players. Players should have a photo and their height listed. Players also play in one position for their career.
- Players can play either goalie, forward, or defense. The latter two are considered "skater" positions. Position affects certain statistics, but otherwise players have the same attributes.
- Players play for one team at a time. A player will play for the same team for a duration of a season. (This is a simplification of reality and further discussed below)
- Seasons are referred to by the years they span, e.g. the current season is the 2019-20 season. Seasons have start and end dates. Each season has a champion, aka the Stanley Cup winner.
- The NHL also uses a salary cap system, with the cap changing each season.
- Teams have unique names consisting of the location they play in and their team name/noun.
- Each team belongs to one conference, either East or West. Conferences are further subdivided into two divisions each (Metropolitan and Atlantic in the East, Pacific and Central in the West). A team belongs to only one division in the conference.
- Each team has a roster for each season. Rosters list player names and positions, as well as their jersey number and salary. While players try to keep the same jersey number their whole career, it can change. Jersey number changes typically occur between seasons.
- Hockey persons can also work for a team as a staff member.
- For this database, staff can be either a head coach or a general manager.
- For staff members, we want to track the hockey person, their title, start date, and end date, as well as which team they worked for during that time.
- Seasons have matches. Matches can be identified by their date and who the home team is. Matches also have a visiting team, home and visiting team scores, a winner, and a loser. Each match can occur in the preseason, regular season, or post-season.
- Players in the roster participate in matches. For each match a player participates in, we want to track match details, player ice time, penalty time, goals scored, points, and assists. For skaters, we want to know how many shots they made on the goal. For goalies, we want to know how many goals they allowed.
- Team standings for each season are determined by points earned in matches. Points are derived from counts of wins, losses, and overtime losses. Games played should also be tracked. Standings are relative rankings within division and conference.
- Player statistics for the season are determined by matches as well. Each player on a roster (player/season/team combination) has stats on the number of games played, ice and penalty time, and their points, goals and assists counts.
- For goalies, we also want to track shutouts, save percentage, total goals allowed, and average goals allowed per game.
- For skaters, we want to know if they were a forward or defenseman and how many shots they made on goal.
- Players win awards. Each award is only given once each season.

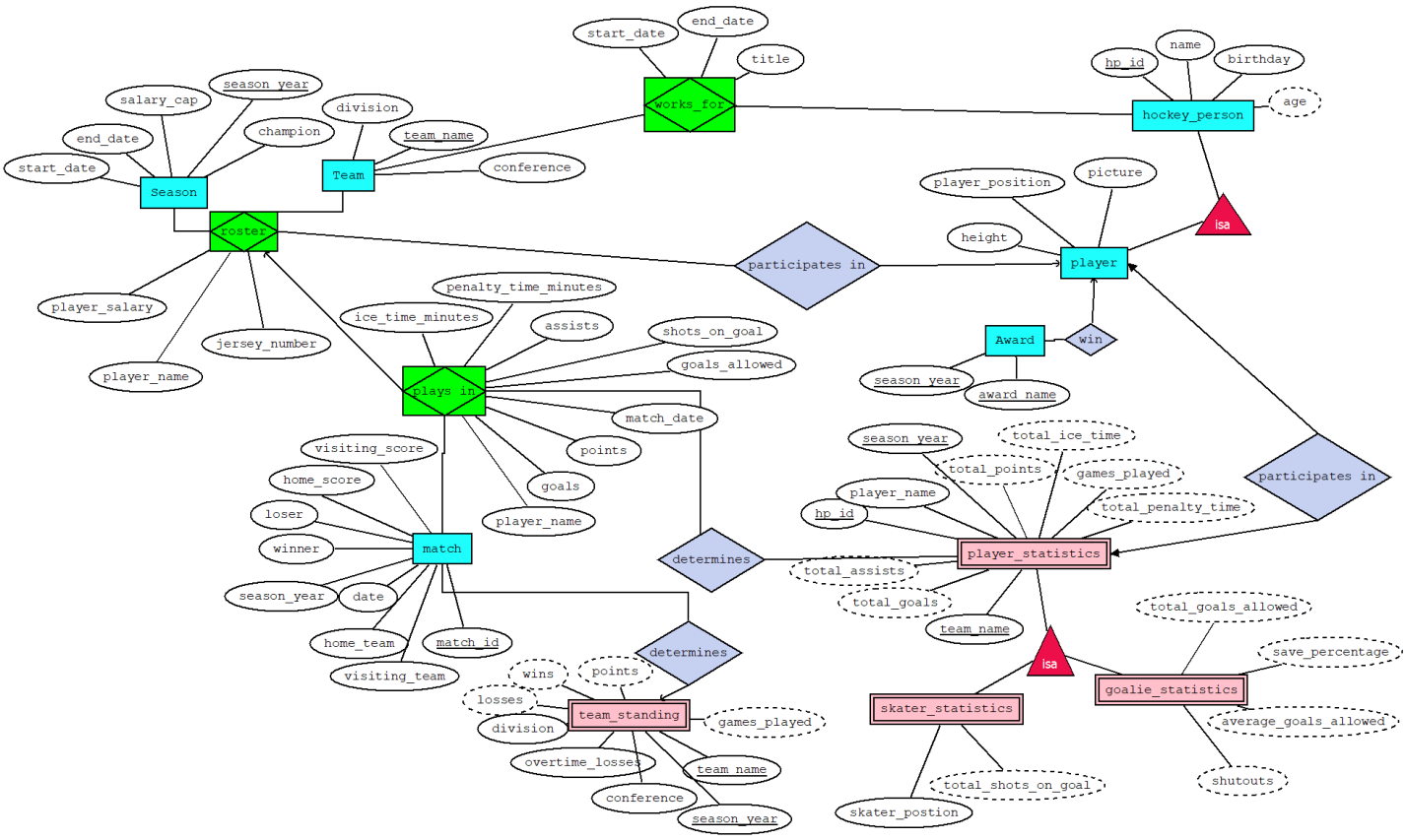

Figure 1: Hockey Domain ERD

1.4 ERD-to-Relations

- hockey_person (hp_id, name, birthday, age)
- Player (hp_id, name, picture, player_position, height_inches)
- team (team_name, conference, division)
- season (season_year, champion, start_date, end_date, salary_cap)
- works_for (hp_id, team_name, staff_name, title, start_date, end_date)
- roster (hp_id, name, season_year, team_name, player_salary, jersey_number)
- award (award_name, hp_id, season_year)
- match (match_id, date, season_year, home_team, visiting_team, winner, loser, home_score, visiting_score)
- plays_in (match_id, season_year, team_name, hp_id, player_name, ice_time_minutes, penalty_time_minutes, points, goals, assists, shots_on_goal, shots_allowed)
- team_standing (team_name, season_year, division, conference, points, games_played, wins, losses, overtime_losses, division, conference)
- player_statistics (hp_id, player_name, season_year, team_name, games_played, total_ice_time, total_penalty_time, total_points, total_goals, total_assists)

- skater_statistics (hp_id, player_name, season_year, team_name, skater_position, total_shots_on_goal)
- goalie_statistics (hp_id, player_name, season_year, team_name, total_goals_allowed, average_goals_allowed, save_percentage, shutouts)

1.5 Explanation and Limitations

The Entity Relationship Diagram above illustrates the entities and relationships in our hockey database (see Figure 1). The hockey person entity encompasses players and staff. Players and staff have different attributes, so players are further modeled as their own entity with an is-a relationship to hockey persons. The staff attributes we chose to model -- such as their title and the time period they held that title -- are really attributes of their time working for a team, so we chose to model the relationship as an associative entity between hockey person and team, instead of modeling a separate staff entity. For simplicity's sake, we intended to only track General Managers and Head Coaches, but there is no restriction on staff titles, so other titles like Assistant Coach are supported. While hockey persons can be uniquely identified by their names and birthdays together, we opted to assign persons unique ids (hp_id) as a more parsimonious primary key.

Seasons are uniquely identified by a seven-character year descriptor, such as 1999-00. Teams similarly have a natural primary key in their team name, which consists of the location and descriptor, e.g. the Toronto Maple Leafs and Montreal Canadiens, rather than just the Maple Leafs and Canadiens. Nicknames like the Leafs and Habs are not supported in this design. A known limitation of this design is that teams occasionally move from one location to another where the same players may play under different team names. For example, the owners of the Atlanta Thrashers sold the team to the Winnipeg based TNSE company and the team became known as the Winnipeg Jets. This is not tracked in our database as this does not happen frequently and may over-complicated the kind of queries our database is designed for. An alternative would have been to model franchises separately from the teams.

Since teams have many players, a player can play for many teams over the course of their career, and play is typically organized by seasons, we modeled that relationship as a roster associative entity whose records or instances would be uniquely identified by a combination of team name, season year, and hockey person ID.

Matches, like hockey persons, can be considered to have a natural composite key, in this case the match date and the home team. As with hockey_person, we chose to assign a unique match ID.

There are six frequently updated entities in our database: the plays_in entity, team_standing entity, the match entity, and the statistics entities (player statistics, goalie statistics, and skater statistics, the latter two of which are subclasses of player statistics). The values contained in each of these entities change with every new match a team plays. Since not every player in a team's roster plays in a match, we modeled 'plays_in' as an associative entity whose keys are inherited from roster and match (See section 1.4). Team standing and player_statistics are necessary weak entities within our database. While tracking statistics is

an important function of the database, these entities do not have any unique identifiers of their own. The values in these entities change frequently as matches are played and goals are scored; many of the attributes are derived. The fast changing nature of hockey necessitates a need for entities where scores are 'held' for the time being before being 'stored'.

We attempted to make our database as detailed as necessary to make it easy to query and extract information from. Given the dynamic nature of hockey, we faced a number of limitations in our attempt to keep our database detailed yet simple. It is common in hockey to have teams trade players in the middle of the season. For our modeling purposes, we assume that players will remain on the same teams for the duration of the season. As a result, players, unlike staff, do not have start dates and end dates given. If we did consider trades as an aspect of players, it would be demonstrated in a separate 'trades' entity. Tables 'player', 'team', 'team_standing', 'match' and 'skater_stats' all have at least 1 column with ENUM constraints. While this is good for consistency, it is a limitation in that if a new value were to be introduced into column 'division' in our 'teams' table, we would have to alter the table to allow for 'division' to accept a new value.

More technical limitations include being unable to enforce data constraints(for example ensuring season end_date is after season start_date), only including the most reputable awards that are awarded to a player in the awards table (Vezina Trophy, Hart Memorial Trophy, etc) and not tracking any in-game awards. Note that the Stanley Cup winning team is not included in the awards table, but appears in the season table as the champion for the appropriate season. Not all possible hockey entities nor statistics are recorded in our database, therefore not all possible statistical queries can be made. Although awards like the Vezina trophy(presented to the top goalie) can be derived from our database, values in the 'awards' entity are inputted manually as some awards (Hart Memorial Trophy) are decided through votes and not purely through player statistics.

Other limitations include player salaries on the 'season' table being capped at 999,999,999.99 per team per season, and not tracking farm team players (players from lower ranks who join the main team). Column widths were capped in cases where value ranges could be expected. For example, goals in a single match are usually in the single digits, and never go as high as 100, so the score columns in match are integer columns with a width of two.

1.6 DDL

```
/* HOCKEY PERSON TABLE */
CREATE TABLE `hockey_person` (
 `hp_id` int(11) NOT NULL AUTO_INCREMENT,
 `name` varchar(45) NOT NULL,
 `birthday` date DEFAULT NULL,
 `age` int(3) DEFAULT NULL,
 PRIMARY KEY (`hp_id`),
 KEY `name_id` (`hp_id`,`name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* PLAYER TABLE */
CREATE TABLE `player` (
 `hp_id` int(11) NOT NULL,
 `name` varchar(45) NOT NULL,
 `player_position` enum('Forward','Defence','Goalie') NOT NULL,
 `height_inches` int(2) DEFAULT NULL,
 `picture` longblob,
 PRIMARY KEY (`hp_id`),
 KEY `name` (`name`),
 KEY `name_id` (`hp_id`,`name`),
 CONSTRAINT `player_to_hp` FOREIGN KEY (`hp_id`) REFERENCES `hockey_person` (`hp_id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* TEAM TABLE */
CREATE TABLE `team` (
 `team_name` varchar(45) NOT NULL,
 `conference` enum('Western','Eastern') NOT NULL,
 `division` enum('Metropolitan','Atlantic','Central','Pacific') NOT NULL,
 PRIMARY KEY (`team_name`),
 KEY `conf` (`conference`),
 KEY `div` (`division`),
 KEY `standings_to_teams` (`team_name`,`division`,`conference`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* TEAM STANDINGS TABLE */
CREATE TABLE `team_standing` (
 `team_name` varchar(45) NOT NULL,
```

```
 `season_year` char(7) NOT NULL,
 `division` enum('Metropolitan','Atlantic','Central','Pacific') DEFAULT NULL,
 `conference` enum('Western','Eastern') DEFAULT NULL,
 `games_played` int(11) DEFAULT NULL,
 `points` int(11) DEFAULT NULL,
 `wins` int(11) DEFAULT NULL,
 `losses` int(11) DEFAULT NULL,
 `overtime_losses` int(11) DEFAULT NULL,
 KEY `standings_to_teams_idx` (`team_name`,`division`,`conference`),
 KEY `standings_to_season_idx` (`season_year`),
 CONSTRAINT `standings_to_season` FOREIGN KEY (`season_year`) REFERENCES `season`
(`season_year`) ON DELETE CASCADE ON UPDATE CASCADE,
 CONSTRAINT `standings_to_teams` FOREIGN KEY (`team_name`, `division`, `conference`)
REFERENCES `team` (`team_name`, `division`, `conference`) ON DELETE CASCADE ON UPDATE
CASCADE
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* SEASON TABLE */
CREATE TABLE `season` (
 `season_year` char(7) NOT NULL,
 `start_date` date DEFAULT NULL,
 `end_date` date DEFAULT NULL,
 `salary_cap` decimal(11,2) DEFAULT NULL,
 `champion` varchar(45) DEFAULT NULL,
 PRIMARY KEY (`season_year`),
 KEY `season_to_winning_team_idx` (`champion`),
 CONSTRAINT `season_to_winning_team` FOREIGN KEY (`champion`) REFERENCES `team`
(`team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci

/* ROSTER TABLE */
CREATE TABLE `roster` (
 `hp_id` int(11) NOT NULL,
 `name` varchar(45) NOT NULL,
 `season_year` char(7) NOT NULL,
 `team_name` varchar(45) NOT NULL,
 `player_salary` decimal(10,2) DEFAULT NULL,
 `jersey_number` int(2) NOT NULL,
 PRIMARY KEY (`hp_id`,`season_year`,`team_name`),
 KEY `roster_to_season_idx` (`season_year`),
 KEY `roster_to_team_idx` (`team_name`),
 KEY `name_idx` (`name`),
 CONSTRAINT `roster_to_player` FOREIGN KEY (`hp_id`) REFERENCES `player` (`hp_id`) ON
DELETE RESTRICT ON UPDATE CASCADE,
```

```
 CONSTRAINT `roster_to_season` FOREIGN KEY (`season_year`) REFERENCES `season`
(`season_year`),
 CONSTRAINT `roster_to_team` FOREIGN KEY (`team_name`) REFERENCES `team` (`team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* WORKS FOR TABLE */
CREATE TABLE `works_for` (
 `hp_id` int(11) NOT NULL,
 `staff_name` varchar(45) NOT NULL,
 `team_name` varchar(45) NOT NULL,
 `start_date` date NOT NULL,
 `end_date` date DEFAULT NULL,
 KEY `works_for_to_hp_idx` (`hp_id`,`staff_name`),
 KEY `works_for_to_team_idx` (`team_name`),
 CONSTRAINT `works_for_to_hp` FOREIGN KEY (`hp_id`, `staff_name`) REFERENCES
`hockey_person` (`hp_id`, `name`),
 CONSTRAINT `works_for_to_team` FOREIGN KEY (`team_name`) REFERENCES `team`
(`team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* PLAYER STATS TABLE*/
CREATE TABLE `player_stats` (
 `hp_id` int(11) NOT NULL,
 `player_name` varchar(45) NOT NULL,
 `season_year` char(7) NOT NULL,
 `team_name` varchar(45) NOT NULL,
 `games_played` int(11) DEFAULT NULL,
 `total_points` int(11) DEFAULT NULL,
 `total_goals` int(11) DEFAULT NULL,
 `total_assists` int(11) DEFAULT NULL,
 `total_ice_time` decimal(8,2) DEFAULT NULL,
 `total_penalty_time` varchar(45) DEFAULT NULL,
 PRIMARY KEY (`hp_id`,`season_year`,`team_name`),
 CONSTRAINT `pstats_to_roster` FOREIGN KEY (`hp_id`, `season_year`, `team_name`)
REFERENCES `roster` (`hp_id`, `season_year`, `team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/*SKATER STATS TABLE*/
CREATE TABLE `skater_stats` (
 `hp_id` int(11) NOT NULL,
 `player_name` varchar(45) NOT NULL,
 `team_name` varchar(45) NOT NULL,
```

```
 `season_year` char(7) NOT NULL,
 `skater_position` enum('Forward','Defence') NOT NULL,
 `total_shots_on_goal` int(11) DEFAULT NULL,
 PRIMARY KEY (`hp_id`,`season_year`,`team_name`),
 CONSTRAINT `skater_stats_to_pstats` FOREIGN KEY (`hp_id`, `season_year`, `team_name`)
REFERENCES `player_stats` (`hp_id`, `season_year`, `team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/*GOALIE STATS TABLE*/
CREATE TABLE `goalie_stats` (
 `hp_id` int(11) NOT NULL,
 `player_name` varchar(45) NOT NULL,
 `team_name` varchar(45) NOT NULL,
 `season_year` char(7) NOT NULL,
 `total_shots_against` int(11) DEFAULT NULL,
 `goals_allowed` int(11) DEFAULT NULL,
 `save_percentage` decimal(4,2) DEFAULT NULL,
 `average_goals_allowed` int(11) DEFAULT NULL,
 `shutouts` int(11) DEFAULT NULL,
 PRIMARY KEY (`hp_id`,`season_year`,`team_name`),
 CONSTRAINT `goalie_stats_to_pstats` FOREIGN KEY (`hp_id`, `season_year`, `team_name`)
REFERENCES `player_stats` (`hp_id`, `season_year`, `team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* MATCH TABLE */
CREATE TABLE `match` (
 `match_id` int(11) NOT NULL AUTO_INCREMENT,
 `match_date` date NOT NULL,
 `home_team` varchar(45) NOT NULL,
 `visiting_team` varchar(45) NOT NULL,
 `season_year` char(7) NOT NULL,
 `home_score` int(2) DEFAULT NULL,
 `visiting_score` int(2) DEFAULT NULL,
 `winner` varchar(45) DEFAULT NULL,
 `loser` varchar(45) DEFAULT NULL,
 `pre_post_reg_season` enum('regular season','postseason','preseason') NOT NULL,
 PRIMARY KEY (`match_id`),
 KEY `match_to_hometeam_idx` (`home_team`),
 KEY `match_to_visitingteam_idx` (`visiting_team`),
 KEY `match_to_season_idx` (`season`),
 KEY `match_to_winner_idx` (`winner`),
 KEY `match_to_loser_idx` (`loser`),
```

```
 KEY `match_date_idx` (`match_id`,`match_date`),
 CONSTRAINT `match_to_hometeam` FOREIGN KEY (`home_team`) REFERENCES `team`
(`team_name`),
 CONSTRAINT `match_to_loser` FOREIGN KEY (`loser`) REFERENCES `team` (`team_name`),
 CONSTRAINT `match_to_season` FOREIGN KEY (`season`) REFERENCES `season`
(`season_year`),
 CONSTRAINT `match_to_visitingteam` FOREIGN KEY (`visiting_team`) REFERENCES `team`
(`team_name`),
 CONSTRAINT `match_to_winner` FOREIGN KEY (`winner`) REFERENCES `team` (`team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci


/* PLAYS IN TABLE */
CREATE TABLE `plays_in` (
 `hp_id` int(11) NOT NULL,
 `player_name` varchar(45) NOT NULL,
 `team_name` varchar(45) NOT NULL,
 `season_year` char(7) NOT NULL,
 `match_id` int(11) NOT NULL,
 `match_date` date NOT NULL,
 `ice_time_minutes` decimal(4,2) NOT NULL,
 `penalty_time_minutes` decimal(4,2) NOT NULL,
 `goals` int(2) NOT NULL,
 `assists` int(11) NOT NULL,
 `points` int(11) NOT NULL,
 `shots_on_goal` int(11) DEFAULT NULL,
 `shots_against` int(11) DEFAULT NULL,
 `goals_allowed` int(11) DEFAULT NULL,
 PRIMARY KEY (`hp_id`,`team_name`,`season_year`,`match_id`),
 KEY `pl_to_match_idx` (`match_id`,`match_date`),
 KEY `pl_to_roster` (`hp_id`,`season_year`,`team_name`),
 CONSTRAINT `pl_to_match` FOREIGN KEY (`match_id`, `match_date`) REFERENCES `match`
(`match_id`, `match_date`),
 CONSTRAINT `pl_to_roster` FOREIGN KEY (`hp_id`, `season_year`, `team_name`) REFERENCES
`roster` (`hp_id`, `season_year`, `team_name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4_0900_ai_ci
```

References

Prewitt, A. (2019, November 14). How Seattle Is Taking a Data-Driven Approach to Building Its New NHL Team. Retrieved from https://www.si.com/nhl/2019/11/14/seattle-hockey-team-analytics-alexandra-mandrycky-ron-francis

Sports Reference LLC. "NHL & WHA Awards and Honors." Hockey Reference, 2020, https://www.hockey-reference.com/awards/

Sports Reference LLC. "2018-19 NHL Goalie Statistics." Hockey Reference, 2019, https://www.hockey-reference.com/leagues/NHL_2019_goalies.html.

Sports Reference LLC. "2018-19 NHL Skater Statistics." Hockey Reference, 2019, https://www.hockey-reference.com/leagues/NHL_2019_skaters.html

Wikipedia. "List of defunct and relocated National Hockey League teams." Wikipedia, 16 January, 2020, https://en.wikipedia.org/wiki/List_of_defunct_and_relocated_National_Hockey_League_teams.