ICE HOCKEY DATABASE SCHEMA DESIGN: FOR NATIONAL TEAM'S BIOMECHANICAL ANALYSIS

Seung-kyo Jin[1,3], Sun-tae Kim[1] Ji-Hyun Jang[2] and Hye-young Kim[2]

Korea Institute of Science and Technology Information (KISTI), Daejeon, Republic of Korea[1]
Korea National Sport University (KNSU), Seoul, Republic of Korea[2]
University of Science & Technology (UST), Daejeon, Republic of Korea[3]

This study presented database schema to manage ice hockey data shown in official record of World Championships. We selected data fields to design database schema considering which fields were contributing to the victory. We also sorted all fields into six tables to reflect the elements of ice hockey games; Game Information, Team Information, Offensice Stats, Defence Stats, Face-off Stats, Time On Ice Stats. This study was a prior research designing database schema in which the analysis of ice hockey official record is used to advance from statistical models to data science techniques. After storing raw data pre-processed into the database, researcher can analyze biomechanically to improve performance.

KEYWORDS: ice hockey, database, schema, data science

INTRODUCTION: Sports analysis is advancing with research based on statistical models. Big data has been applied in all industries as a resource that can increase their efficiency. Demand for collecting, managing, and analysing data generated by sports stakeholders is growing. Big data analysis in sports has been divided into sports business and for improving of performance. Alamar (2013) presented that teams such as the Oakland A's, Tampa Bay Rays, and San Antonio Spurs have embraced the use of analytics, and all three clubs, though they are in small markets and so have limited resources, have seen tremendous success in part because of the information edge gained by their analytics programs. Big data analysis seeks to understand the meaning that can be extracted from data points. Collecting and managing data required highly experienced experts and physical resources, both of which are cost and effort expensive. Thus, winter sports national team in Korea have not been well equipped with data management infrastructure like big sports clubs all over the world. This study represented the data warehouse storing big data in ice hockey games among winter sports events, concerned with data pre-processing. Raw data to analyse the database schema was analysed on the Datanest platform developed by Korea Institute of Science and Technology Information (KISTI).
In Korea, Cho (2012) insisted that no real examples of big data infrastructure or applications have been found to provide sufficient evidence for the effectiveness and strategic value of big data. Park and Lee (2013) emphasized that sports public data should also strengthen its role in fulfilling the right of the public to know and in offering a chance to accumulate new intellectual property. Swartz (2017) emphasized Ice hockey is a difficult sport for analytics and have plentiful data. These difficulties and challenges are looking as opportunities. There have already been attempts to harness big data. Mehrasa (2018) proposed generic deep learning model, one of machine learning methodologies, that learns powerful representations of player trajectories. Korea National Sport University (KNSU) was conducting a project to unify the management of the data generated from various winter sports events in Korea. KNSU's research was meaningful to lay the foundation for biomechanical data management in winter sports for Korea national team.

METHODS: We designed database schema as a basis of stochastic methods and data science including machine learning, deep learning and reinforcement learning etc. Roith and Magel (2014) surveyed main variables contributing to the wins based on the ice hockey World Championships official result sheets. 1) 23 fields composed database schema. Among them, 22 fields were relevant with the main variables selected based on the results of the

World Championships game records. International Ice Hockey Federation(IIHF) placed ranking of the participating countries for the World Championships using those 22 fields announced contributing to the victory in the Roith and Magel's study. One more field was selected as an ID. Official game summary sheet summarizes game information where, when, which teams play, goalkeeper records, game statistics, periodic summary and team statistics. Researcher can extract those fields from the game summary sheet. 2) We normalized whole data fields except ID to get the database schema and divided into six tables as official record can be divided into six parts: Game Information, Team Information, Offensice Stats, Defence Stats, Face-off Stats, Time On Ice Stats.

RESULTS: Every field and its explanation to design database schema is on the table 1. ID field was not in raw data, but it is created to identify each game uniquely. Group, Year, Game, Win_lose, Country, means team name, were on the official game summary sheet. Score, Save, SOG, PIM, TPP, PPG and SHG were on the team statistics table in the same sheet. Score_diff was derived from the gap between each team's scores. FOW, FOL, FO_net and FOW% were derived from the game statistics tables. EQ was derived from score minus PPG and SHG. SG%, SVS%, PP% represents proportion of score with shots on goal, saves from shots on goal, goals with scored under the power play. TPP represents time for power play using integer as minutes and floating points as seconds. Contrary to this, TPPinMin represents Power Play Time proportionally.

Table 1: Fields of database schema

Field	Explanation	Field	Explanation
Group	Group	Score	Scores
Year	Year	Score_diff	Scoring difference
Game	Game order	Save	Saves
Country	Team name	ID	Unique ID
Win_lose	Win or lose	SOG	Shots on goal
SG%	Scored out of shots	PIM	Penalties in minutes
TPP	Time on power play	PPG	Power play goals
SHG	Shorthanded goals	FOL	Face-off lose
FOW	Face-off win	FO_net	Face-offs net
EQ	Goals under same number of players	FOW%	Percentage of Face-offs wins
SVS%	Save percentage on Shots on goals	PP%	Percentage of goals scored in power play
TPPinMin	TPP in minutes	-	-

As presented in Figure 1, fields are classified as 6 sorts of tables. Game Information, Team Information, Defensive Stats, and Face-off Stats that show probability taking the puck when referee drops the puck to start of restart the game. Face-off affects shot on goals and game result. Time On Ice Stats is concerning with different playing time caused by the imbalance of players because of the foul. Tables are connected through ID. ID, Country, Group have VARCHAR type, described by character. Win_lose uses BOOLEAN type to distinguish win as true, lose as false. Score_diff, FOW, FOL, FO_net, Year, Game, PIM, Save, EQ, SHG, PPG have INT type so that represent integer. FOW%, SVS%, SG%, PP% are FLOAT type to show floating point number. ID field is a primary key of the Game Information table which specifies row uniquely in a table. ID is also foreign key in the other tables. All of the tables are connected with ID field. If database is designed followed the schema on Figure 1, it will be scalable to add another information represented in raw data.

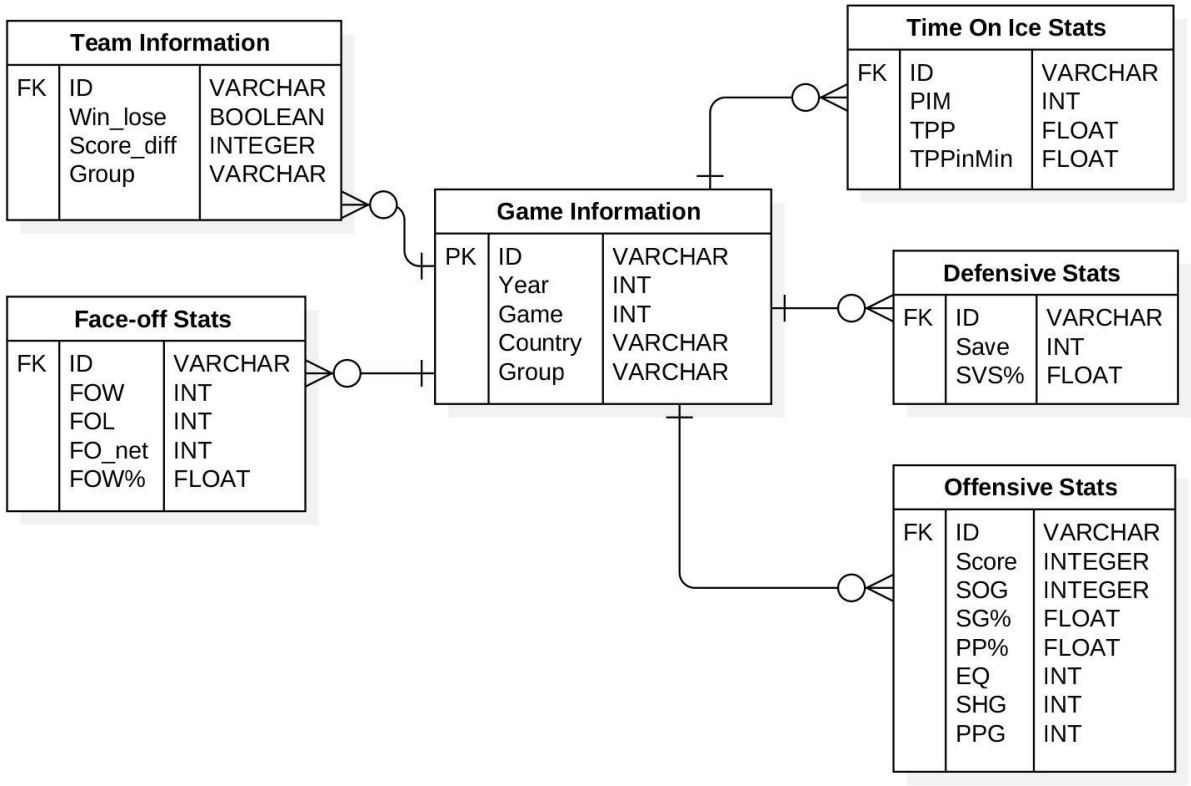

Figure 1. Database schema for ice hockey.

DISCUSSION: This study is important as it designs a database schema to systematically store, retrieve and extract data for the national ice hockey team, which received relatively less attention than professional teams did. Pre-processing the raw data is one of the difficult process in data science. This research reduces the entry barriers to data science in winter sports, which can help researchers and coaching staff to set the team's plan more biomechanically effective. There will be three main possible follow-up studies. First, data analysts can use this database to analyse big data by increasing the number of samples. Second, researchers can expand the range of the dataset to train model from the National Hockey League games as well as the World Championships games used in the prior study. Third, researchers can produce services so that coaches who are unfamiliar with data handling or data science methodologies can use the database to analyse data easily. However, it is needed to discuss more about the service for the unfamiliar users. Because the purpose of a data analysis is to improve the performance in the actual competition, it is necessary to continue debating which data to store for the data analysts and coaching staffs.

CONCLUSION: This research represented data fields, tables and database schema to store the data from the ice hockey official records. For inserting raw data into the database, how and what to pre-process the raw data is important and difficult. Through using the database ice hockey data analysts and coaching staffs are not going to have much burden on pre-processing the dataset, which is taking long time in the data science process. Managing data using the database lead to further research. National Hockey League or other sports clubs use not only statistics methodologies but also machine learning and deep learning especially reinforcement learning to analyse each team biomechanically. Through storing and managing the data, Korean national hockey team can apply data science.

REFERENCES

Benjamin C. Alamar (2013). *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers.* New York, NY: Columbus University Press

Cho, J. (2012). Utilization and Prospect of Sport Big Data, *Journal of Measurement and Evaluation in Physical Education Studies*, 14, 1-11.

Swartz, T. B. (2017). Hockey Analytics. *In Wiley StatsRef: Statistics Reference Online* (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels). doi:10.1002/9781118 445112.stat07965

Park, S., Lee, J. (2013). The Future Utility of Sports Public Big Data and the Plan for Its Further Development, *Journal of Sport and Leisure Studies*, 54, 539-546.

Roith and Magel (2014). An analysis of factors contributing to win in the National Hockey League, *International journal of sports science* 4(3): 84-90

Nazanin Mehrasa (2018). Deep Learning of Player Trajectory Representations for Team Activity Analysis, *2018 MIT Sloan Sports Analytics Conference*

ACKNOWLEDGEMENTS: This research was supported by Sports Science Convergence Technology Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2014M3C1B1034028).