**PLAYERS - FIFA 22**

**PART 3: PROJECT REPORT**

**Instructor:**

Heather Matheson

**COMP 3380 Group 29:**

Mohamed Hammam

Camila Martinez

Dhvani Thakkar

**1. SUMMARY OF THE DATA SET**

Our dataset is called Players FIFA 22 . It provides statistics about players, clubs, and national teams for the Career Mode of FIFA 22 and the World Cup. This dataset allows comparisons of attributes and performance from players or teams, which is valuable for both fans and researchers interested in football. The reason our group chose this database is because all three of us are big fans of the FIFA World cup that occurs every 4 years . Hence, exploring this dataset related to FIFA players, made working on this project even more exciting and easier for testing our queries since we were already familiar with some metrics and data related to the players.

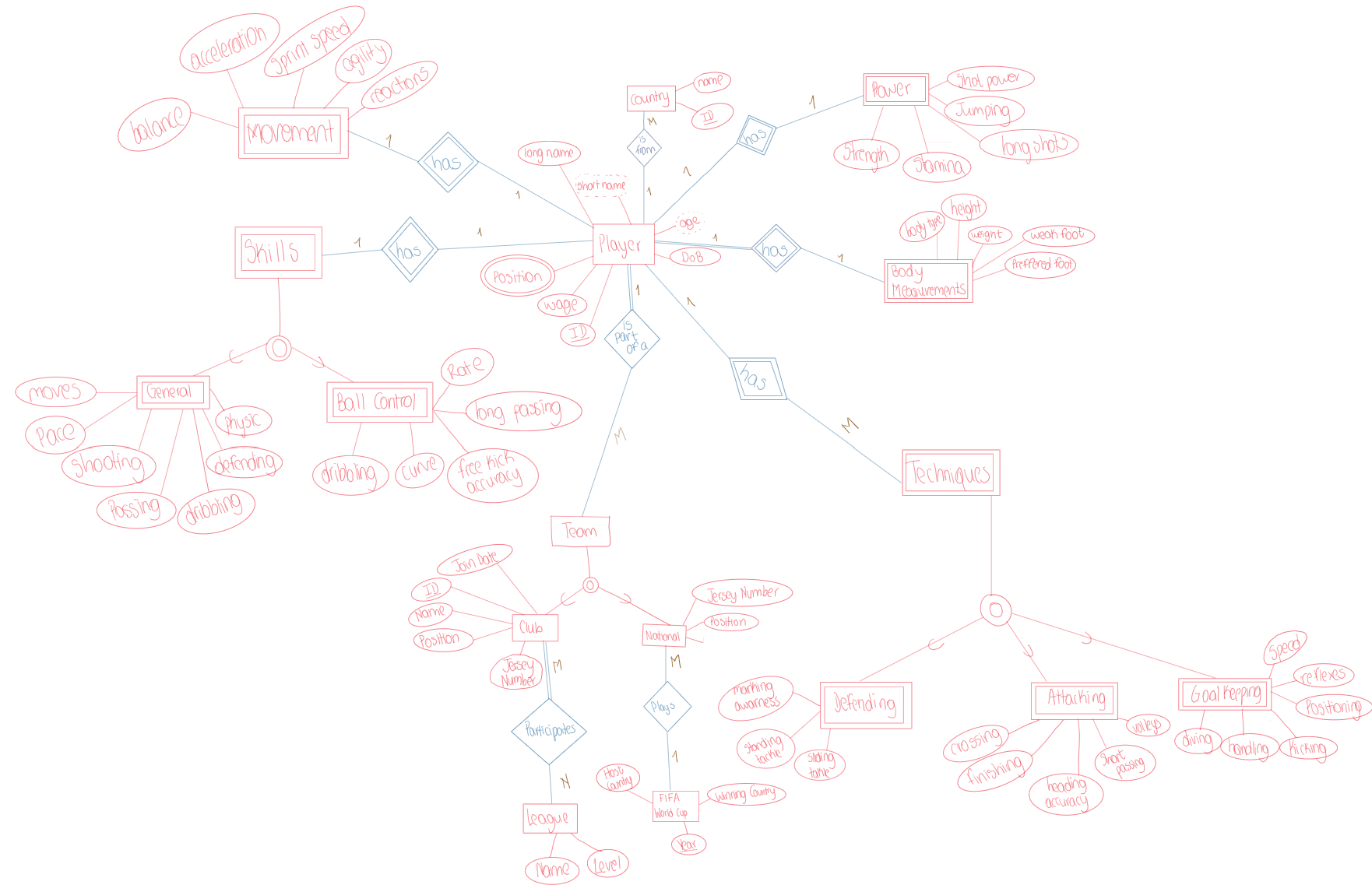
This dataset consists of:

* 16 tables
* 269387 insertion queries
* 19240 players from 164 countries
* 703 clubs

The CSV file for the FIFA 22 Player dataset was downloaded from kaggle.com and the author is Stefano Leone: [FIFA 22 Players Dataset](https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?rvi=1). The CSV file for the FIFA World Cup was

also downloaded from kaggle.com and the author is Andre Becklas: [FIFA World Cup Dataset](https://www.kaggle.com/datasets/abecklas/fifa-world-cup)

The following is the ER Diagram of our FIFA 22 Players Data set:

**The ER Diagram**

**2. DISCUSSION OF THE DATA MODEL**

The data model was created to show all the attributes and qualities that make up a FIFA player. In order to guarantee data integrity, minimize redundancy, and enable efficient data processing, it was divided into many tables. Every table has a distinct entity and attributes specific to that entity. There are several elements of a player's FIFA profile that were taken into consideration when creating the tables such as the "General Skills", "Ball Control Skills”, "Movement Skills", etc. Each of these denote a distinct aspect of a player's skills which makes it simpler to update, retrieve, and change data, and makes the structure more understandable and organized.

In addition to that, dealing with the multivalued attribute "positions" in the "Player" entity might have been one of the major difficulties we faced. Since the characteristics in a relational database should be atomic, handling the many-to-many relationships between players and positions required us to create a separate table for it.

The data model easily fitted well into the relational database. Foreign keys are used to represent the relationships between each object, which are each represented as a table. The data can be updated and queried efficiently because of this structure.

Based on the actual work with the data, various changes to the original data model may have been made. For instance, more tables or relationships were required, and some properties needed to be changed like the multi-valued attribute in table Player and also division of tables that was done after normalization. The purpose of any modifications was to enhance the database's effectiveness, consistency, and usability.

**3. ALTERNATIVE MODEL**

Although there are always other approaches to data modeling, we believe that the selected model appears to successfully represent the essential characteristics of a FIFA player. If the performance is a problem, modifications may be made based on requirements or limitations.

With the work completed, this model appears to be a wise selection. It seems to offer a thorough representation of the data and to make operations more efficient.

**4. INTERESTING QUERIES**

**Query 1**

All the players that play for the specific given club and for the specific given country. Both the country name and club name can be replaced to get the output of players accordingly.

Returns: The player’s ID, the player’s long name, the club's name, and the country’s name

The reason why we find this query interesting is because this can be super helpful especially if you want to find the players that are part of a certain club and are from the same country.

**Query 2**

All the players that have played in a national team but not in a club.

Returns: The player’s ID and player’s long name

The reason why we find this query interesting is because this query can help us find all the players that don't participate in the leagues since they are not part of a certain club but do play for the national team during the FIFA World Cup which occurs once in 4 years. This can also help clubs choose a new player if needed since that player is not part of a club and joins it.

**Query 3**

All the players who are in a certain club and are older than a certain age in descending order.

Returns: The player’s ID, the player’s long name, the club's name, and the country’s name

The reason why we find this query interesting is because this can be super when we are trying to filter out the older players in a certain group since the skills and the body measurements of a player might change after a certain age and it can affect their performance too so this way if a club wants to remove players who are older than a certain age, we can easily filter them out using this query.

**Query 4**

All the players that are in the specific given club whose contract ends in that specific year. The year should be after the year 2022 since the database is from 2022.

Returns: The player’s ID, the player’s long name, the club's name, and the country’s name

The reason why we find this query interesting is because this can super useful is because a player’s contract needs to be renewed or valid before they play for their club and using this query it can be simple for the club to find players whose contracts are expiring in a specific year so that the renewal process can be started immediately or if they have time till the contract expires.

**5. RELATIONAL DATABASE AND ALTERNATIVE DATABASES**

For this dataset, it is required to use a relational database since this is a really big dataset since it consists of 19,000+ rows, and to view and modify the dataset it would be extremely difficult and time-consuming if we try to do it manually or on a CSV file. Using a relational model, we can assure that the modifications can be implemented faster and the data won’t be lost in case of a crash of the computer or the program. The advantage of mainly using a database would be that it follows ACID (Atomicity, Consistency, Isolation, and Durability) which gives us a guarantee that our data is consistent and reliable as well as that the transactions performed are smooth.

Graph databases would not have been a good way to display this dataset. This is because it would have been time-consuming to convert the CSV files into cypher code and slow to run them onto Neo4j. The visual representation of graph databases would have been super useful for this dataset since the Players attribute is like the main attribute connecting to all other attributes; however, the program could take a long time to load the data since we have at least 19000 players in the dataset along with various related attributes. Another disadvantage of this would be that Neo4j does not save the previous cypher codes written to view data in graph or table format.

These interesting queries mentioned above would have been more difficult to recreate in another database modeling system since we would not achieve the level of precision and accuracy in the other data model. The SQL queries for the above queries are super simple and human-readable too but if the same query was written in cypher code it would have been complex to both implement and read the code. We would be able to create more interesting queries cypher code but again the readability and complexity of the code would have been a big issue for the user.

**6. FUTURE USE OF THIS DATASET IN COMP 3380**

One of the best advantages is that most people are familiar with some players, clubs, or countries’ performances from the FIFA World Cup or Leagues which would make understanding some query results and testing more straightforward. One of the issues that we faced during class was getting a result but not being able to know if it was the right answer since going through all of the data is tedious work. However, while testing our queries, it was easy to know if we got good results since we had an idea of what to expect. For this reason, we believe this database would be good for sample queries and practice for COMP3380.

This database contains some inconsistencies by having some rows with missing values, or data format errors. Students might focus on developing processes to clean and improve the quality of data within the dataset. Another reason why this dataset would be of good use for future COMP 3380 students is because it is a large dataset so working on it would be a good way to implement the skills that are learnt throughout the course.

**7. METHODOLOGY OF WORK**

Our methodology of work was collaboration and division of tasks. We split up tasks among the team according to our strengths in the different topics, making sure everyone had their piece to focus on. Then, we would meet regularly to explain how we did things, propose ideas, and get feedback from everyone.

Mohammed focused primarily on the creation of the database, and populating the database with data. For adding the records, he developed a Java program that accepted the CSV, parsed each row, and returned an appropriate insert query.

Camila focused on the development of the front-end interface. For this, she designed a Java program that connects to the database, formats the output properly, displays appropriate error messages, and prevents SQL injection by using PrepareStatement.

Dhvani focused on the creation of queries and testing that they would return the proper results. Most queries allow input from users. To achieve this, she used different methods learned during classes like aggregate functions such as: GROUP BY, ORDER BY, and WITH.

**8. CONCLUSION**

Lastly, we enjoyed working on all three parts of this project as it helped us have a deeper understanding of the course material and put into practice the skills we learned in class on our data set FIFA 22 Players to modify the data, normalize it, and to make a front end interface for it.