Report Group 28

Daniel Mai - Axel Le - Hien Nguyen

1. Summary of data

1.1. Why was it chosen?

- At first, the dataset for our group is from "2016 Olympics in Rio de Janeiro | Kaggle", which is missing a connection from the event table to an athlete table. After considering carefully, with Adam's permission, we have decided to use a very similar dataset from "Tokyo 2020 Olympics | Kaggle", which will make a slight change to the ER-ERR model and relational model. For your convenience, we will include the new one at the end of this report.
- Since the Olympics is the biggest sporting event, the dataset of the Olympics will contain many interesting things to explore. Moreover, it is easy to understand and make queries for all three of us.

1.2. What does it consist of?

The original dataset consists of 5 csv file: athletes, coaches, medals,medals total, technical officials

1.3. How large is it? (File size, number of records)

- athletes:	2.11MB,	11629 rows and 14 columns
- coaches:	60.21KB,	392 rows and 9 columns
- medals:	441.91KB,	2401 rows and 12 columns
- medals total:	2.44KB,	93 rows and 7 columns
- technical officials:	137.76KB,	956 rows and 8 columns

* *Note:*

- There are lots of null values in column birth_country, discipline, gender.
- For discipline, because it is important to link between the discipline table and the athlete table, I have to fill in the correct discipline for each athlete based on the discipline code (which has no missing value). For gender and birth_country, I keep it null as it is (since for some query that asks about the country that athletes represent for, we have the country column which has no missing value).
- There are also some duplicates (name) in technical, which I have pre-process carefully.

2. Discussion of data model

- We broke down the data into 9 tables, which are: athletes, country, discipline, coach, medals, wins, summary, officials, and supports.
- Since country → countryCode, discipline → disciplineCode, medalCode → medalType, we have three tables which are country, discipline and medals.

- The WINS table actually is the medals table from the original dataset, but now, it lacks the medalType, which is in the medals table. You can have it by linking the WINS table with the MEDALS table.
- Supports and officials table come from the technical officials in the original dataset. The Supports table is a table that shows the official name and their discipline accordingly while the officials table will contain all information of those officials.
- Summary is just the medal totals table in the original data. I decided to get rid of the total medal at the end of the table so that we can have a query to do that for us.

3. Summary of the database

- athletes: 6 rows, 11629 rows
- coach: 7 rows, 392 rows
- country: 2 rows, 206 rows
- discipline: 2 rows, 46 rows
- medal: 2 rows, 3 rows
- officials: 5 rows, 875 rows
- summary: 5 rows, 93 rows
- supports: 2 rows, 956 rows
- wins: 5 rows, 2401 rows

4. List of queries

4.1. Easy

4.1.1. Return all the countries available in the database.

4.1.2. Return all the discipline in the database.

4.1.3. Return athletes participate in the same discipline X and are from a country Y, including their gender and birthdate. X and Y are input parameters.

4.1.4. Show all provided events in a discipline.

4.1.5. Show the total number of medals of the top X countries, ordered by the rank announced by the end of the Olympic.

4.1.6. Summarize the number of athletes each country sends

4.1.7. Summarize the number of athlete per discipline

4.2. Medium

4.2.1. Find the oldest/youngest athlete that participates in a given discipline. Return his name, age, and the discipline he plays.

4.2.2. Retrieve top X women who achieve the most models, list the discipline they play and the country they come from.

4.2.3. Find a country's youngest athlete, show the discipline he plays and the type of medal he obtained.

4.2.4. Categorize athletes of each country into discipline, summarize the total of athletes in each entry.

4.2.5. Summarize the number of medals of each athlete in a given combination of country and discipline.

* *Note:* since not all countries gain medals, some combination of country and discipline could result in an empty table. Please give it another try if it happens.

One tested pair of parameter is ("People's Republic of China"- "Badminton")

4.2.6. Name the athletes that won the highest medal in the discipline he plays and his coach, including their country.

4.2.7. Retrieve the discipline that is most popular.

There are two options:

+ popular within a country, based on the number of athletes play it
+ popular worldwide, based on the number of countries play it

4.3. Hard

4.3.1. List all countries that participate in all available disciplines.

* *Note:* This query execution time is fairly long. ~1 minute

4.3.2. In an given country X that has medals in every event of a discipline Y, name all of their athletes who won the medals, the events and medal type.

* *Note:* This query has limited possible parameters. For testing purpose, an example is ("People's Republic of China"- "Badminton") or ("People's Republic of China"- "Table Tennis")

5. Summary of the interface

5.1. The interface was created with java, without any external libraries.

5.2. The zip file needed to run the interface contains:

- ExeIntFac.java, MyDatabase.java
- ExeIntFac.class, MyDatabase.class
- db_project.db
- sqlite-jdbc-3.36.0.3.jar

5.3. How to use the interface:

- Open command prompt of Windows, type:

```
java -cp ".;./sqlite-jdbc-3.36.0.3.jar" ExeIntFac
```

- Type h to see all the possible commands.
- Type the command exactly as it is, with a comma and a space between each component.
- With commands that require arguments (specified by single apostrophes), type the arguments without single apostrophes.

  For example: write command *events, 'dName'* as *events, Swimming*
- With command *athBirth, max/min, 'dName'*, only type either max or min, not both.

Relational Model:

Athletes (**name**, shortname, gender, birth_date, country_code, discipline_code) FK country_code REF Country, FK discipline_code REF Discipline

Country (**countryCode**, countryName)

- countryCode → countryName

Discipline (**disciplineCode**, Discipline)

- disciplineCode → Discipline

Coach(**name**, gender, birth_date, country_code, discipline, function) FK country_code REF Country, FK discipline_code REF Discipline

Medals(**medal_code**, medal_type)

- Medal_code → medal_type

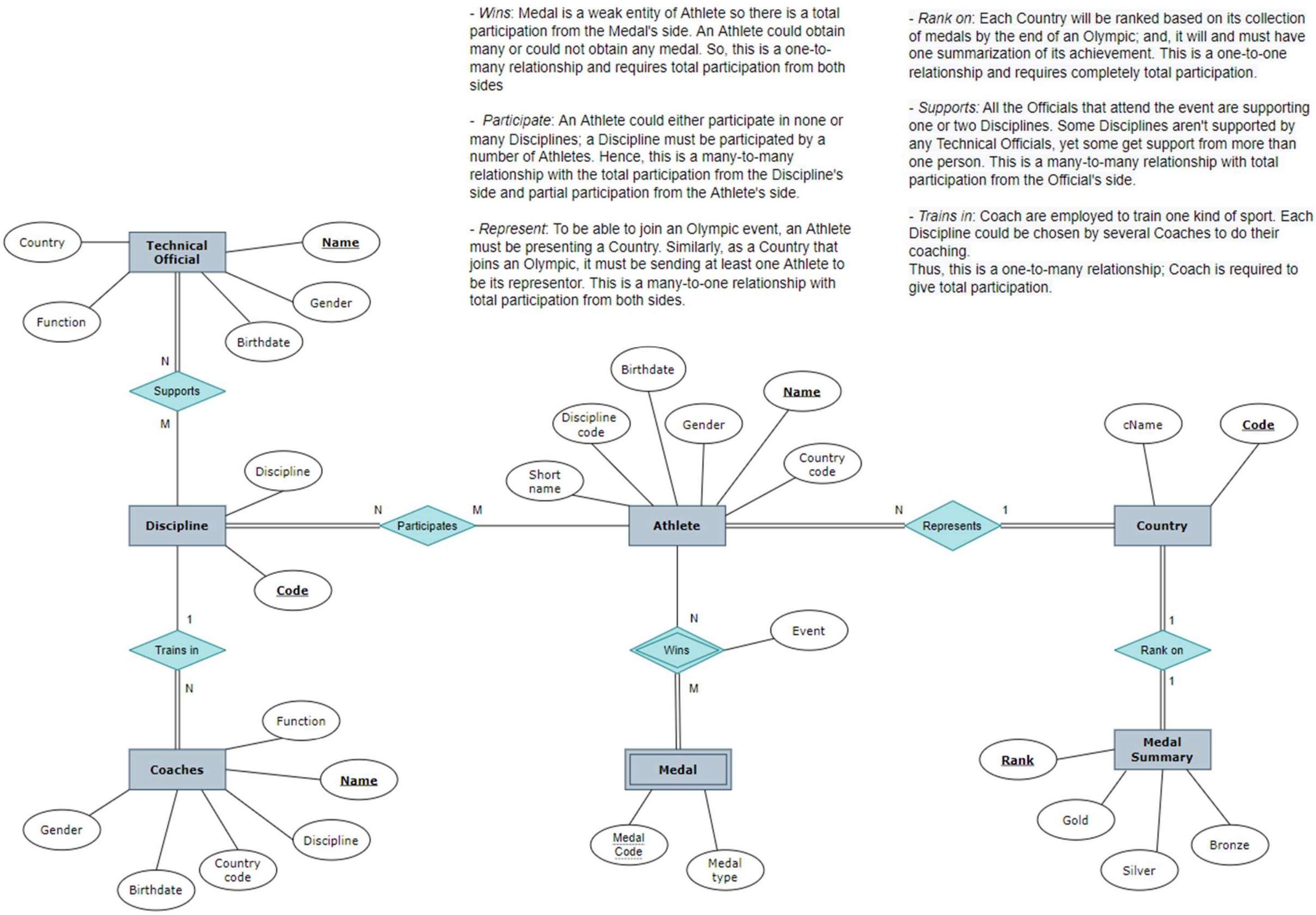
Wins (**athlete_name, medal_code, event**, country_code, discipline_code)
FK athlete_name REF Athletes(name),
FK medal_code REF Medals,
FK country_code REF Country,
FK discipline_code REF Discipline

Summary (**Rank**, country_code, Gold Medal, Silver Medal, Bronze Medal) FK country_code REF Country

Officials (**name**, gender, birth_date, country, function) FK country REF Country(country_name)

Supports (**name, discipline**) FK name REF Technicals, FK discipline REF Discipline(discipline)

ER Model:



- *Wins*: Medal is a weak entity of Athlete so there is a total participation from the Medal's side. An Athlete could obtain many or could not obtain any medal. So, this is a one-to-many relationship and requires total participation from both sides

- *Participate*: An Athlete could either participate in none or many Disciplines; a Discipline must be participated by a number of Athletes. Hence, this is a many-to-many relationship with the total participation from the Discipline's side and partial participation from the Athlete's side.

- *Represent*: To be able to join an Olympic event, an Athlete must be presenting a Country. Similarly, as a Country that joins an Olympic, it must be sending at least one Athlete to be its representor. This is a many-to-one relationship with total participation from both sides.

- *Rank on*: Each Country will be ranked based on its collection of medals by the end of an Olympic; and, it will and must have one summarization of its achievement. This is a one-to-one relationship and requires completely total participation.

- *Supports:* All the Officials that attend the event are supporting one or two Disciplines. Some Disciplines aren't supported by any Technical Officials, yet some get support from more than one person. This is a many-to-many relationship with total participation from the Official's side.

- *Trains in*: Coach are employed to train one kind of sport. Each Discipline could be chosen by several Coaches to do their coaching.
Thus, this is a one-to-many relationship; Coach is required to give total participation.