Project Report

Team 5: Ari Freireich, Sai Gangineni, Arifun Nabi, Tao Lin, Haaris Malik
INST 327

Introduction

For our Relational Database Project this semester, our team is interested in creating a database pertaining to the largest sporting event in the world, The Olympic Games, and particularly the Running Events within the category of Athletics. Specifically, we hope to collate together all of the data about the participating countries, the individual competitions, the World/Olympic records, and the athletes themselves who have won Olympic medals and thereby distinguished themselves at the top of the field of running in the entire world. Our ultimate objective with this database is to provide a central location where users can find all info about Olympic Medal Winners from around the world for all running events over several years. By combining all of this information, we hope to allow for various comparisons to be made such as the difference between countries' performances during the Olympics or the performance of a particular athlete between separate Olympic events. By having this type of extensive and detailed dataset, users will be able to follow along easier and understand more about one of the greatest traditions of our planet that has persisted for thousands of years.

The database will be created using publicly available data from the Internet about the many Olympic Games as well as the participating countries. The various tables within the database will be created based on each entity such as Countries, Athletes, Sporting Events, etc. and will be populated with data from multiple different websites to fulfill our goal of centralizing this information. The first modern Olympics was held in 1896 in Athens, however, our research will focus on the Olympics held from 2000 to 2020 to focus on the athletes that our current generation would be most familiar with. Although there are many participants every time the Olympic Games are held, our project will focus on only the medal winners from each sporting event as well as their countries. The sporting events themselves will include both Men’s and Women’s running events from the Summer Olympics, however, we will not be including other events within the category of Athletics like Javelin Throw, Shot Put, Hurdles, etc. Along with this, we will not be including other international sporting competitions such as the Asian Games,

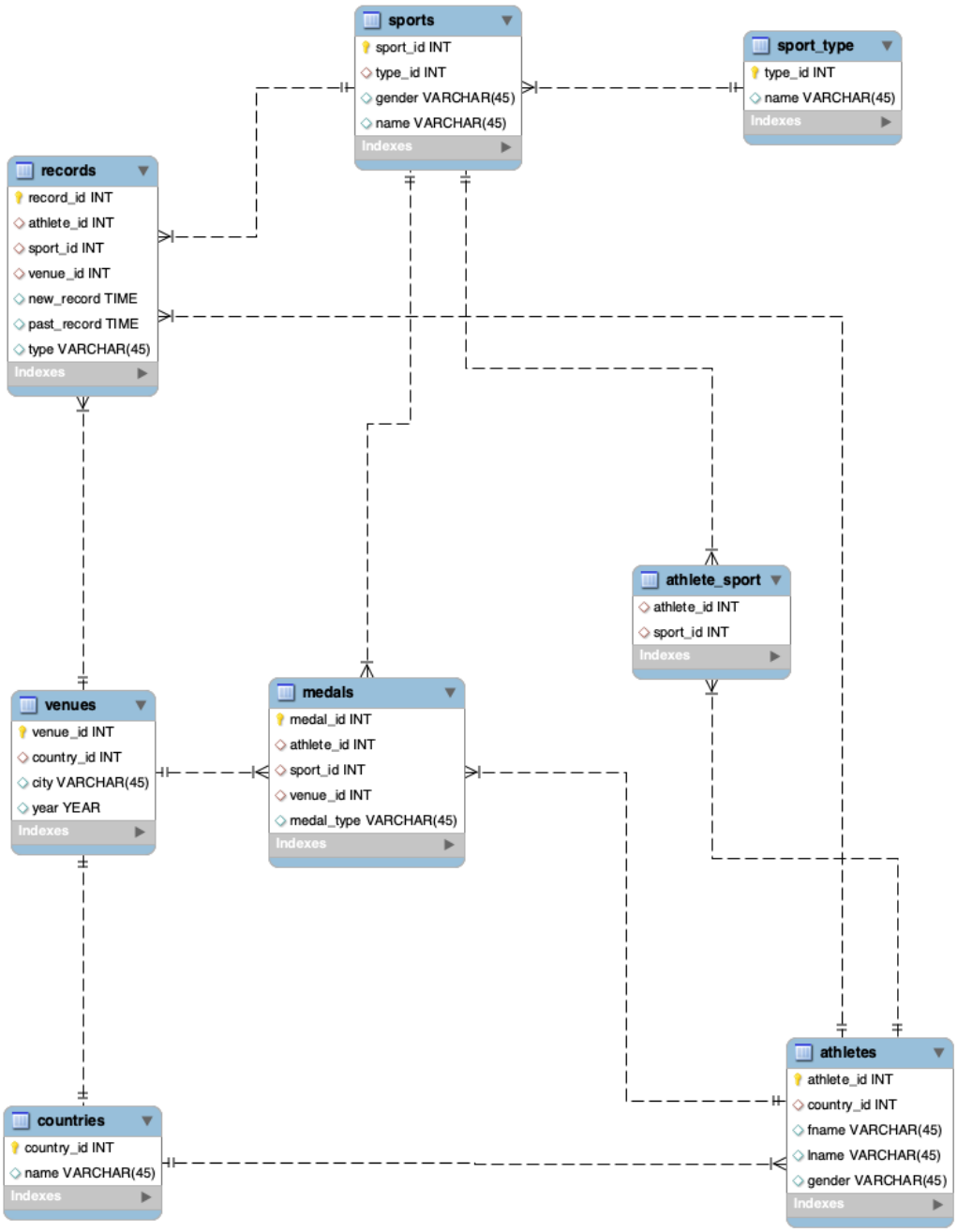
Commonwealth Games, etc. even though those competitions hold their fair share of historical records and feats.

Although our data is pulled from online websites, the structure of our database itself will be uniquely tailored for our project. This means that certain specific data about sporting events such as the exact running times of events will be excluded while more details in other aspects will be added such as specifics about participating countries and athletes. Much of our database design is structured around the idea of making comparisons and trends to understand how countries, athletes, sporting events, and even the Olympics itself has evolved over the years. The information that is easily available online and what most of the public are shown are always recent statistics due to the four year gap between each Olympic Games, so we wanted to offer the possibility of examining more of the history so users can be more informed about their favorite sports or athletes.

The users who would use this database can be the people who organize the Olympics, the people who manage leaderboards/scoreboards, or people who want to see the results, athletes, or sport information to then publish online for the world to see. Even the general public would be interested in this project as they would be able to see for themselves the comparisons and trends that can be drawn from the data. Due to the large gap between the Olympic Games, the public really only know about recent medal winners or their favorite athletes while watching the games, but with our database, they will be able to be better informed about the history behind their favorite events and learn more information whether it is about countries, sports, or athletes.

Database Description

- **Logical Design**



- **Physical Database**

- Data relating to identifying the participating countries in The Olympic Games. This table would allow for users to view basic information about a country and would be labeled as **Country**. This table is connected in a one-to-many relationship with the Athletes table and a one-to-one relationship with the Venues table. The columns within the table are:
  - country_id (Primary Key)
  - name (String)
- Data relating to individual athletes and labeled as **Athletes**. This table would include all the athletes who have won an Olympic Medal within the last 20 years within the Athletics category as well as specific information about them such as their country and name. This table is connected in a one-to-many relationship with the Medals and Records tables. The columns within the table are:
  - athlete_id (Primary Key)
  - country_id (Matches Country table)
  - fname (String)
  - lname (String)
  - gender (String)
- Data relating to the individual sporting events within the Athletics Category and labeled as **Sports**. This would include all the specific running events within the Athletics category such as 100m, 400m, 10,000m, etc. This table is connected in a one-to-many relationship with the Medals and Records tables. The columns within the table are:
  - sport_id (Primary Key)
  - type_id (Matches Sport_Type table)
  - gender (String)
  - name (String)
- Data relating to the category of running event, specifically between the options of sprints, mid-distance running, long-distance running, race walking, and marathons. This table is called **Sport_Type** and is connected in a one-to-many relationship with the Sports table. The columns within the table are:
  - type_id (Primary Key)

        - name (String)
    - Join (Linking) table connecting the Athletes and Sports tables. This table is called **Athlete_Sports** and would connect the primary keys of both tables in a many-to-many relationship. The columns within the table are:
        - athlete_id (Primary Key - Matches to Athlete table)
        - sport_id (Matches Sports table)
    - Data relating to the specific Olympic Games conducted every four years and labeled as **Venue**. This would allow for users to view the foundational information about the Host Countries that have held the Olympic Games over the years. This table is connected in a one-to-many relationship with the Medals and Records tables. The columns within the table are:
        - venue_id (Primary Key)
        - country_id (Matches Country table)
        - city (String)
        - year (Int)
    - Data relating to the World and Olympic Records broken by various athletes for different sporting events over the years. This would be labeled as **Records** and would allow for users to view information about the feats of athletic prowess by some of the greatest in the world with information including the sporting event, athlete information, as well as the past record and the new record accomplished. The columns within the table are:
        - record_id (Primary Key)
        - athlete_id (Matches Athlete table)
        - sport_id (Matches Sports table)
        - venue_id (Matches Venue table)
        - past_record (Time)
        - new_record (Time)
        - type (String)
    - Data relating to the Olympic medals won by athletes. This table would be labeled as **Medals** and would allow for users to view the information of medal winners

for different events for each year of the Olympics. The columns within the table are:

- medal_id (Primary Key)
- athlete_id (Matches Athlete table)
- sport_id (Matches Sports table)
- venue_id (Matches Venue table)
- medal_type (String)

- **Sample Data**
  - Our data was pulled from an actual real-life olympic database located in the website, https://www.olympiandatabase.com/. This is a publicly available dataset with extensive information for every Olympics since 1952 with separate Tables/Datasets for Sporting Events, Athletes, Countries, and Medals. We used the information about specifically the Running events within the sport of Athletics as well as the specific athletes that won a medal for those events between the years of 2000 and 2020. This data was inputted manually into an Excel worksheet due to lack of options to download it as a csv file. After editing and formatting the dataset using Excel, we directly imported it to Workbench into the database tables using the Data Wizard.

- **Views/Queries**

View Name	Req. A	Req. B	Req. C	Req. D	Req. E
Query 1	**X**	**X**	**X**		**X**
Query 2	**X**	**X**	**X**		
Query 3	**X**	**X**	**X**		
Query 4	**X**	**X**			
Query 5	**X**	**X**		**X**	
Query 6	**X**	**X**	**X**		
Query 7	**X**	**X**	**X**		

*****Queries found at bottom of report and in separate file**

Changes from Original Design

Our initial design was to create an all-inclusive Olympic Games database that would include all data from all Olympic Games, both Summer and Winter Olympics, from the first recorded year in the online public database which is 1952 to now. But, after assessing the feasibility of creating such a large dataset by ourselves, we changed our idea to only include **Olympic Medal Winners** since they are the athletes at the top of their fields. However, the dataset was still too large for us to add, especially without an easy csv file available for download so we decreased it further down to only the category of Athletics within the Summer Olympics, specifically the **Running Events** (100 m, 200 m, 400 m, 800 m, 1500 m, 5000 m, 10000 m, Marathon, 20 km Race Walk, 50 km Race Walk). We also added data about **World and Olympic Records** broken and limited the data to the Olympic Games from **2000 to 2020** (last five games) as those are the games within our generation.

Lessons Learned

Database ethics had zero impact on our project since much of our information could be found through a simple google search and there was no privacy, copyright, or data misuse concerns. All we did was make it easier to find the information a user would want rather than them having to search for these results through a few more clicks. Furthermore, our database really only shows the result of already recorded public data from the Olympics and compiles it into something easier to find. We aren't really storing any information about the user that tries to assess our database. Our data also was sourced from a website that we believe to not have stolen, forged, or obtained illegally about the Olympics. They also listed it publicly so we believe we also haven't illegally taken their data since we aren't using it for profits. This leads us to believe that we didn't violate any database ethics while the topic of the project itself being a common and worldwide phenomenon means that ethics doesn't play a role with our database.

We learned a lot during the course of this project about MySQL, Workbench, and even Excel. While coming up with our initial ideas at the beginning of the semester, we really loved the idea of the Olympics because its a popular sporting event and it would have been such a great thing to have all the information about this international event at our fingertips for easy browsing through

a database. But, we unfortunately were not able to carry through with the idea. The lack of a downloadable public csv file online, with the only available ones being too large for any of us to feasibly download and later format on Excel without the computer hanging up, meant that we had to scrap our idea and change it. This was actually a great lesson for us as we got to go through several meetings just refining and redesigning which was great practice for real-life situations where things might not work the way we want it to. Along with this, we got a lot of practice with normalization in a real situation as we edited our initial ERD diagram and changed a lot of tables around when normalizing. Finally, our best practice was with the use of Workbench as a tool to add, delete, and modify tables, databases, and views along with manipulating queries to answer different research questions.

Potential Future Work

For potential future work, we would have our database include all data about the Olympics like our initial data. This would mean data from all years in which the Olympics have been held, all events, both summer and winter, and other international sporting events (World Championships, etc.) With the future for databases expanding, there can be more data sets to set up an entirely new and wide variety of databases. Along with this, we would like to have highlighted more visible trends in performance during the Olympic Games between athletes, sports, and countries. Although our current database allows us to see comparisons between athletes and countries for the running categories in the last few Olympics Games, we would definitely be interested in seeing the same type of trends when taking all of the hundreds of sporting events into account.

Appendix

1.
```
CREATE OR REPLACE VIEW foreign_medal_winners AS
SELECT CONCAT(a.fname, ' ', a.lname) AS athlete,
        (SELECT c.name
                FROM countries c
      WHERE c.country_id = a.country_id) AS country_name,
        COUNT(m.medal_id) AS num_medals
```

FROM medals m

JOIN athletes a ON m.athlete_id = a.athlete_id

WHERE (SELECT c.name FROM countries c WHERE c.country_id = a.country_id) != 'USA'

GROUP BY a.athlete_id

ORDER BY num_medals DESC;

2. CREATE OR REPLACE VIEW most_female_medals AS

SELECT c.name AS 'Country', a.gender AS 'Gender',

COUNT(m.medal_type) as 'Total Medals'

FROM countries c

INNER JOIN athletes a ON c.country_id = a.country_id

INNER JOIN medals m ON a.athlete_id = m.athlete_id

WHERE a.gender = "Female"

GROUP BY name

ORDER BY count(m.medal_type) DESC;

3. CREATE OR REPLACE VIEW ethiopian_medals AS

SELECT c.name AS country, s.name AS sport, COUNT(m.sport_id) AS num_medals, st.name AS sport_type

FROM medals m

JOIN athletes a USING(athlete_id)

JOIN countries c USING(country_id)

JOIN sports s ON m.sport_id = s.sport_id

JOIN sport_type st USING(type_id)

WHERE c.name = "Ethiopia"

GROUP BY m.sport_id

ORDER BY num_medals DESC;

4. CREATE OR REPLACE VIEW olympic_records AS

SELECT CONCAT(a.fname, " ", a.lname) AS athlete_name,

a.gender, r.new_record, s.name, r.type

FROM records r

JOIN athletes a ON a.athlete_id = r.athlete_id

```
        JOIN sports s ON r.sport_id = s.sport_id
    WHERE type != 'World'
    ORDER BY a.lname;
```

5.
```
    CREATE OR REPLACE VIEW athlete_categories AS
    SELECT DISTINCT CONCAT(a.fname, ' ', a.lname) AS athlete,
    st.name AS sport_category
    FROM athletes a
        JOIN athlete_sport atp USING(athlete_id)
        JOIN sports s ON atp.sport_id = s.sport_id
        JOIN sport_type st USING(type_id)
    ORDER BY athlete;
```

6.
```
    CREATE OR REPLACE VIEW top_five_countries AS
    SELECT c.name AS country, COUNT(m.medal_id) AS num_gold
    FROM medals m
        JOIN athletes a USING(athlete_id)
        JOIN countries c USING(country_id)
    WHERE medal_type = "Gold"
    GROUP BY country
    ORDER BY num_gold DESC
    LIMIT 5;
```

7.
```
    CREATE OR REPLACE VIEW kenya_golds AS
    SELECT c.name AS country, s.name AS sport, COUNT(m.medal_id) AS num_gold
    FROM medals m
        JOIN athletes a USING(athlete_id)
        JOIN countries c USING(country_id)
        JOIN sports s ON m.sport_id = s.sport_id
        JOIN sport_type st USING(type_id)
    WHERE m.medal_type = "Gold" AND c.name = "Kenya"
    GROUP BY s.name;
```