Estadísticas de los Juegos Olímpicos en Neo4j

ZuoHeng Dai
Santiago Costa
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay

Resumen

El presente documento expone el desarrollo del proyecto final de la materia Bases de Datos No Relacionales. Este consiste en el diseño y creación de una base de datos de grafos utilizando la herramienta Neo4j a partir de datos de los Juegos Olímpicos, con el fin de analizarlos y relevar ciertas estadísticas sobre los atletas y sus equipos. Asimismo, se pretende evaluar el rendimiento de la herramienta Neo4j y la ejecución de algoritmos sobre grafos.

I. Introducción

Las bases de datos de grafos son un tipo de base de datos no relacionales que modelan la información empleando grafos, donde los nodos representan entidades de la realidad, las aristas representan relaciones que vinculan entidades y tienen una dirección. Asimismo, se introduce el concepto de etiqueta para agrupar nodos en conjuntos. Este tipo de modelo, permite lidiar con relaciones complejas, y el hecho de no tener un esquema definido permite trabajar de forma efectiva con un gran volumen de datos [1] [2]. Neo4j es una plataforma que permite la implementación de bases de datos sobre grafos de manera eficiente, y provee un lenguaje muy potente de consultas, *Cypher*, así como una librería de algoritmos sobre grafos denominada *Graph Data Science* (GDS).

El desarrollo de este proyecto consiste en la utilización de la herramienta Neo4j para el diseño e implementación de una base de datos de grafos para el análisis y gestión de datos sobre los Juegos Olímpicos. En particular, estos datos describen la participación de los distintos atletas en la historia de las olimpiadas desde Atenas 1896 hasta Río 2016, indicando los resultados de cada competición.

Los objetivos que se plantean son los siguientes:

- Analizar los datos y realizar consultas en el lenguaje *Cypher* de Neo4j para responder a ciertas preguntas sobre el dominio en cuestión, de forma de obtener estadísticas relevantes sobre atletas y sus equipos.
- Ejecutar algoritmos sobre grafos de la librería *Graph Data Science* para la detección de nodos importantes y comunidades.
- Evaluar el rendimiento de la herramienta Neo4j para una base con un gran volumen de datos, atendiendo al tiempo de ejecución de las consultas y algoritmos.

La base de datos cuenta con más de 100.000 nodos y 300.000 relaciones, donde las consultas efectuadas sobre la misma fueron exitosas, permitiendo extraer información y estadísticas enriquecedoras sobre atletas y sus países. Asimismo, los algoritmos empleados dilucidan acerca de la composición del grafo exponiendo aquellos nodos con mayor relevancia y detectando comunidades de nodos que compartan similitudes según ciertos criterios. Estos resultados permiten concluir el gran poder que presenta la herramienta Neo4j para el manejo de grandes volúmenes de datos, así como su versatilidad y flexibilidad para el análisis de los mismos.

El resto del documento se organiza de la siguiente forma. En la Sección II se presenta el desarrollo del proyecto, detallando la obtención de los datos, el pre procesamiento de los mismos, el diseño de la base y la carga de dichos datos. Luego, en la Sección III se presentan las consultas efectuadas a la base y sus análisis correspondientes, así como ejecuciones de algunos algoritmos sobre grafos. Por último, en la Sección IV se presentan las conclusiones finales del proyecto elaborado y algunas consignas sobre trabajo a futuro.

II. Desarrollo

II-A. Obtención de los datos

Los datos empleados para el desarrollo y construcción de la base fueron extraídos del sitio web *Kaggle* [3], una plataforma de ciencia de datos de acceso público, que comprende, entre otros aspectos, distintos conjuntos de datos sobre temáticas diversas. Como bien se aludió previamente, estos datos conforman un *dataset* histórico de los Juegos Olímpicos incluyendo los juegos desde Atenas 1896 hasta Río de Janeiro 2016. La composición de los datos consiste de dos archivos de extensión *.csv* que fueron descargados del sitio con el fin de efectuar un análisis y pre procesamiento de manera local. El primero, *athlete_events.csv* (41.5 MB), contiene mas de 270.000 filas y quince columnas, donde cada fila corresponde a un atleta individual compitiendo en un evento o disciplina olímpica específica. Del atleta se conoce sus características físicas, su equipo o país que representa, la edición de los juegos en la que participó y qué medalla obtuvo en dicho evento. El segundo archivo, *noc_regions.csv* (3.6 KB), contiene 230 filas y tres columnas, donde cada fila corresponde a un código de identificación único de tres letras denominado NOC (*National Olympic Committee*) junto con el nombre del país asociado que dicho código representa.

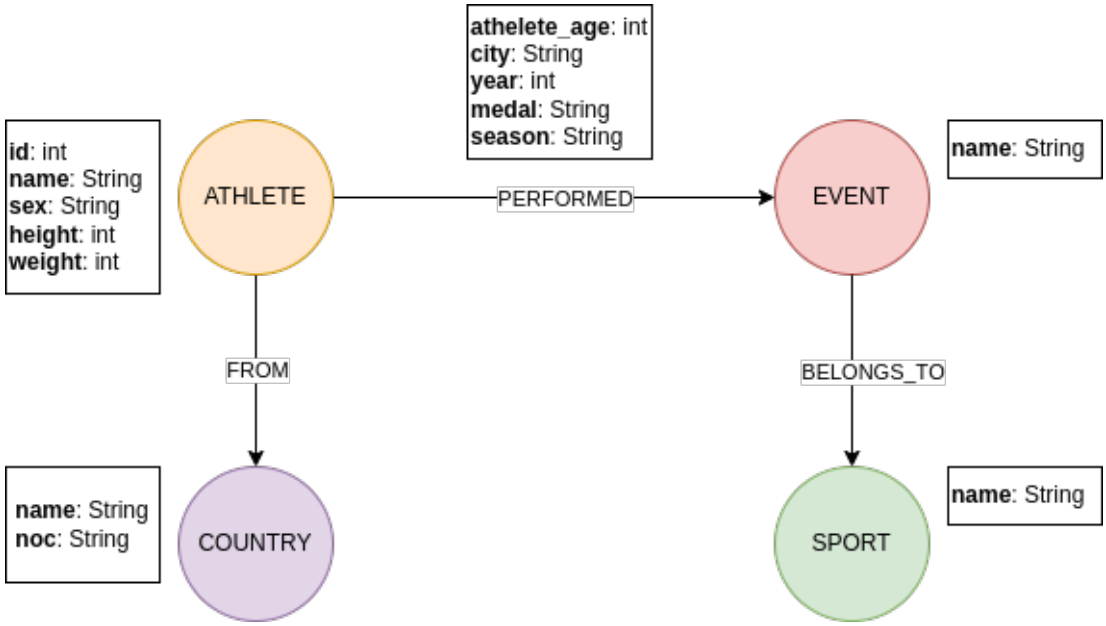

Figura 1: Diagrama de diseño de la base de datos de grafo.

II-B. Análisis y pre procesamiento de los datos

Como antesala al diseño y elaboración de la base, se efectuó un extenso proceso de estudio y análisis de los datos con el fin de lograr una comprensión en profundidad de los mismos y eliminar aquellos que no se consideren pertinentes a los efectos del proyecto. Esta tarea fue llevada a cabo sobre los datos *athlete_events*, utilizando el lenguaje Python y la librería *pandas* para su manipulación y análisis. El resultado de este pre procesamiento se almacena en un nuevo archivo *olympic_data.csv*. Las modificaciones realizadas se describen a continuación.

II-B1. Columnas: Se decidió eliminar la columna *Games* dado que la información que representa corresponde a la combinación de los valores de las columnas *Year* y *Season*, por lo cual no aporta valor y es redundante.

Por otra parte, la columna *Team* refleja el equipo de competición, comúnmente el nombre del país. Esto, sin embargo, no era así en los primeros años de los juegos, pudiendo existir más de un equipo perteneciente a un mismo país, aunque en la columna *NOC* asociada el código es el mismo para todos los equipos de un mismo país. Dado que para el propósito de este estudio lo que interesa es el nombre del país como equipo, se decidió tomar el código de país NOC y luego utilizando el *dataset noc_regions*, se mapea dicho código al nombre del país, logrando así agrupar todos los resultados de dichos equipos como un solo país.

II-B2. Valores nulos: Examinando los datos, se notó también la presencia de valores nulos en algunas columnas. Dado que Neo4j no almacena este tipo de valores, se tomaron ciertas acciones para enmendar esta contingencia. Los valores nulos de la columna *Medal* indican que el atleta no ganó ninguna medalla en el evento deportivo que desempeñó. Por tanto, se reemplaza cada valor nulo por el *string* “No Medal”.

Asimismo, se notó la presencia de varios valores nulos en las columnas *Weight* y *Height*, especialmente en atletas que participaron en las primeras ediciones de los juegos. En base a esto, se tomo la decisión de eliminar aquellas filas de atletas que contengan un valor nulo en algunas de esas dos columnas, exeptuando aquellos atletas hayan obtenido al menos una medalla, pues este último dato es relevante para el análisis y consultas futuras sobre los datos.

Por último, para aquellas columnas *Age*, *Weight* y *Height* que aún sigan teniendo valores nulos, estos serán reemplazados por el valor -1.

II-C. Diseño de la base

Para el diseño de la base de datos se tuvo en cuenta las consultas que se querían realizar a futuro. En este sentido, se identificaron las distintas entidades de la realidad que se pretendían modelar y cómo estas se relacionan. En la figura 1 se observa un diagrama del grafo con los nodos y relaciones correspondientes.

Particularmente y en vista de los objetivos definidos, se consideró modelar como nodos a los atletas (ATHLETE), sus países (COUNTRY), los eventos que desempeñó (EVENT) y los distintos deportes olímpicos existentes (SPORT). Estos nodos interactúan mediante las tres relaciones siguientes:

- La relación FROM representa el país de un atleta.
- La relación PERFORMED representa la participación de un atleta en un evento específico. Las propiedades de la relación indican en qué edición de los juegos desempeñó ese evento y qué medalla obtuvo.
- La relación BELONGS_TO representa el deporte del cual cada evento forma parte.

II-D. Carga de los datos a la base

Para la carga de datos a Neo4j se intentó, primeramente, emplear el comando `LOAD CSV` de *Cypher*, pero dado el gran volumen de datos en cuestión, la carga requería de bastante tiempo y parecía no finalizar. Razón por la que se optó por cargar los datos desde Python utilizando las librerías *neo4j* y *pandas*. No obstante, la carga requería de un tiempo considerable, por lo que se decidió agilizar aún más el proceso separando en distintos archivos de extensión *.csv* los datos a cargar. Es decir, a partir del *olympic_data.csv* se crearon archivos subconjunto de este último reduciendo así la información de carga. De esta forma, no se tiene que recorrer el archivo entero para, por ejemplo, cargar solamente los nodos que representan los países, pues estos están contenidos en un archivo más pequeño que contienen únicamente todos los países. Estos archivos fueron creados a través del *script generate_loads_csv.py*. Vale aclarar que para este trabajo se utilizó la versión de escritorio de Neo4j (*Neo4j Desktop*).

II-E. Disponibilidad de los datos

Todos los archivos de datos y *scripts* de Python se encuentran disponibles en el repositorio de GitLab del proyecto [4]. A continuación se describen brevemente los *scripts* utilizados

- ***preprocessing.py***: realiza el pre procesamiento de los datos y genera el *olympic_data.csv*.
- ***load_nodes.py***: efectúa la carga de los nodos a la base en Neo4j.
- ***load_relationships.py***: efectúa la carga de relaciones a la base en Neo4j.

III. Experimentación

III-A. Descripción de la base de datos

La base de datos cuenta con los siguientes números de nodos y relaciones.

Nodos:

- Athlete: 107.057
- Country: 227
- Event: 760
- Sport: 66

Total: 108.110 nodos

Relaciones:

- FROM: 108.625
- PERFORMED: 216.415
- BELONGS_TO: 759

Total: 325.799 relaciones

En la figura 2 se visualiza un ejemplo de un grafo representativo de la base diseñada, donde en color rojo se representa a los nodos Sport, en celeste nodos Event, en violeta nodos Athlete y en anaranjado nodos Country.

III-B. Consultas

En esta sección se realizan consultas de interés de forma de relevar ciertas estadísticas relevantes sobre los juegos olímpicos, en especial sobre atletas y sus países así como generalidades de las distintas disciplinas olímpicas.

III-B1. Atletas con más medallas: Se busca responder a la pregunta de cuáles fueron los atletas que obtuvieron más medallas en toda la historia de los juegos. Vale aclarar que un mismo atleta puede participar en más de una edición de los juegos, y por ende la cantidad de medallas corresponde a la suma de las medallas ganadas en cada edición. En el listado 1 se visualiza el código de la consulta.

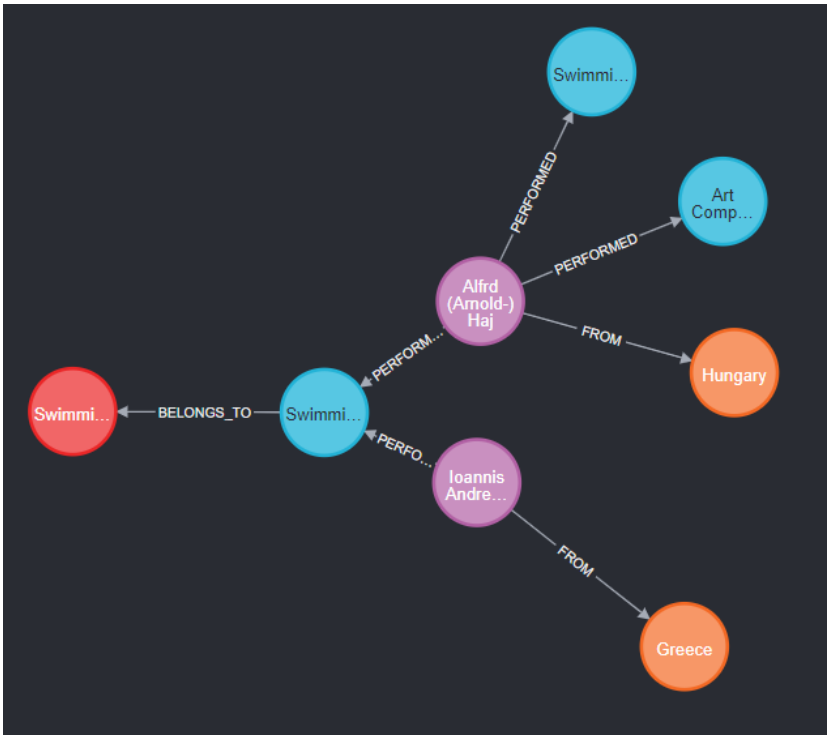

Figura 2: Grafo de ejemplo.

Listado 1 Consulta de atletas con más medallas

```
MATCH (a:Athlete)-[r:PERFORMED]->(e:Event)
WHERE r.medal IN ['Gold', 'Silver', 'Bronze']
RETURN a.name AS Athlete,
   COUNT(r) AS TotalMedals,
   COUNT(CASE r.medal WHEN 'Gold' THEN 1 END) AS GoldMedals,
   COUNT(CASE r.medal WHEN 'Silver' THEN 1 END) AS SilverMedals,
   COUNT(CASE r.medal WHEN 'Bronze' THEN 1 END) AS BronzeMedals
ORDER BY TotalMedals DESC, GoldMedals DESC, SilverMedals DESC, BronzeMedals DESC
```

En la tabla I se observan los diez atletas con más medallas ganadas y su distribución entre oro, plata y bronce. En el primer lugar del *ranking* se encuentra el gran medallista estadounidense Michael Phelps con un total de 28 medallas ganadas y 23 de ellas, de oro.

III-B2. Países con más medallas de oro: Esta consulta es similar a la anterior con la diferencia de que se sitúa el foco en los países, de modo de evaluar su competitividad en base a sus triunfos en las distintas disciplinas olímpicas.

Inicialmente, se efectuó una consulta para determinar qué países han ganado mayor cantidad de medallas de oro en la historia de los juegos. En el listado 2 se expone la consulta empleada.

Athlete	TotalMedals	GoldMedals	SilveMedals	BronzeMedals
Michael Fred Phelps, II	28	23	3	2
Larysa Semenivna Latynina (Diriy-)	18	9	5	4
Nikolay Yefimovich Andrianox	15	7	5	3
Ole Einar Birndalen	13	8	4	1
Borys Anfiyanoyych Shakhlin	13	7	4	2
Edoardo Mangiarotti	13	6	5	2
Takashi Qno	13	5	4	4
Paavo Johannes, Nurmi	12	9	3	0
Birgit Fischer-Schmidt	12	8	4	0
Jennifer Elisabeth ”Jenn”Thompson (-Sumpelik)	12	8	3	1

Tabla I: Atletas con más medallas

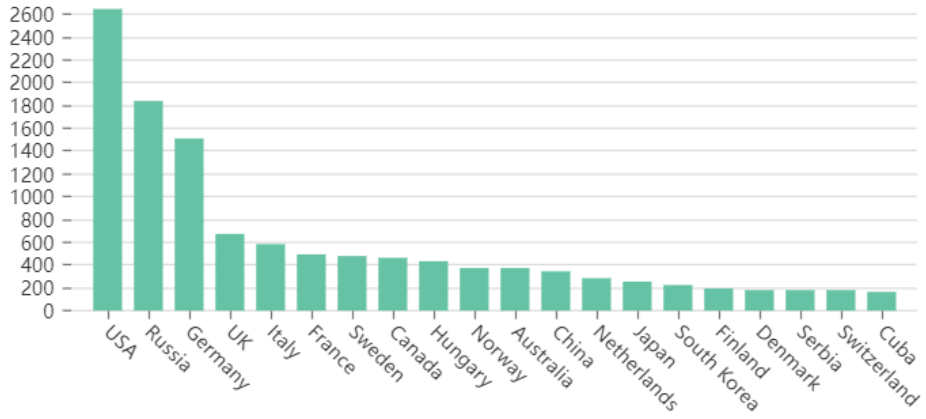

Figura 3: Paises con más medallas de oro.

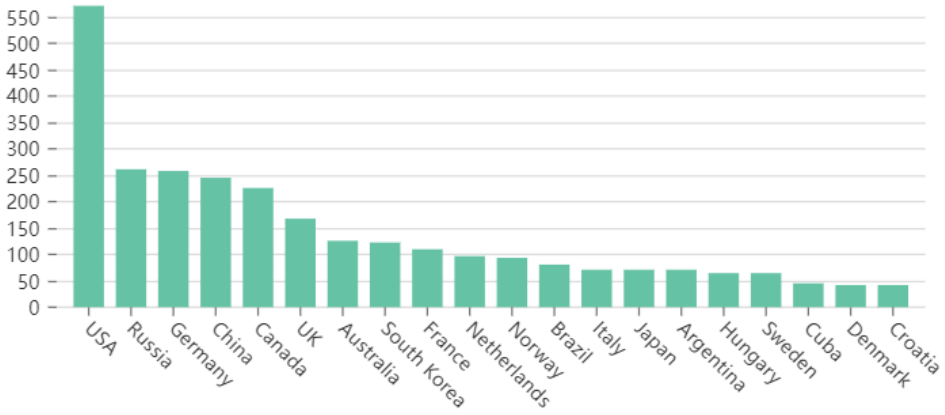

Figura 4: Países con más medallas de oro desde el año 2000.

Listado 2 Consulta de países con más medallas de oro.

```
MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED {medal: "Gold"}]->(e:Event)
RETURN c.name AS CountryName, COUNT(p) AS TotalGoldMedals
ORDER BY TotalGoldMedals DESC
```

En la figura 3 se observa una gráfica con el *ranking* de los veinte países con más medallas de oro obtenidas.

Luego, se decidió explorar cómo variando los valores de ciertas propiedades de la relación PERFORMED, podría afectar significativamente los resultados. En primer lugar, se agregó la condición `WHERE p.year >2000` para limitar los resultados a los eventos que ocurrieron después del año 2000. Esto permitió analizar los países que han sido más exitosos en términos de medallas de oro en el pasado más reciente. La consulta realizada se dispone en el listado 3.

Listado 3 Consulta de países con más medallas de oro según restricción de año.

```
MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED {medal: "Gold"}]->(e:Event)
WHERE p.year > 2000
RETURN c.name AS CountryName, COUNT(p) AS TotalGoldMedals
ORDER BY TotalGoldMedals DESC
```

Observando la figura 4, se aprecia una diferencia significativa con los resultados de la consulta anterior. Si bien los primeros tres países continúan manteniendo su posición, se puede notar que China paso de estar en la doceava posición a estar cuarto en el *ranking* junto con una cantidad de medallas muy similar a la de Rusia y Alemania. Por tanto, se deduce un aumento de la competitividad por parte de China en las ediciones de los juegos más recientes. Asimismo, Canadá y Australia presentan un compartimiento muy similar.

Por ultimo, se realiza la misma consulta agregando la condición `season: 'Winter'` para restringir los resultados a eventos de temporada de invierno. El listado 4 contiene la consulta correspondiente.

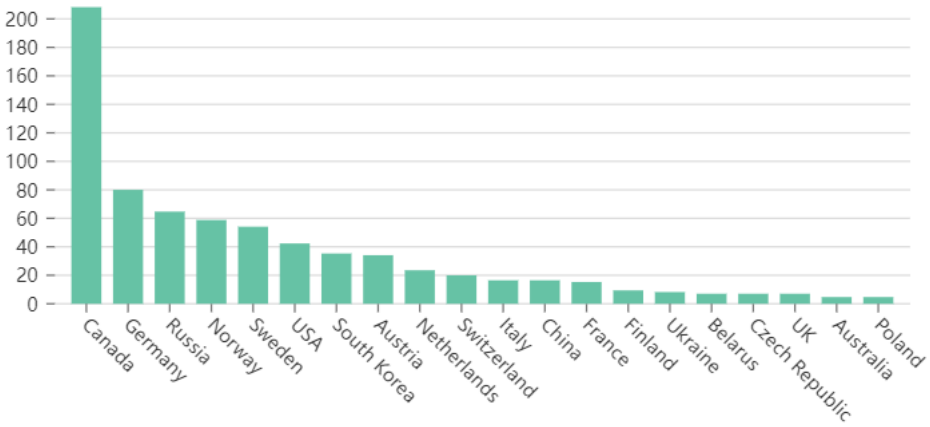

Figura 5: Países con más medallas de oro desde el año 2000 y temporada de invierno.

Listado 4 Consulta de países con más medallas de oro según restricción de año y temporada.

```
MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED {medal: "Gold", season: "Winter"}]->(e:Event)
WHERE p.year > 2000
RETURN c.name AS CountryName, COUNT(p) AS TotalGoldMedals
ORDER BY TotalGoldMedals DESC
```

Visualizando la gráfica de la figura 5, se nota un cambio radical en el *ranking* de los países, siendo Canadá el que lidera con amplia diferencia. Por otra parte, EE.UU. desciende a la sexta posición mientras que países como Noruega y Suecia ascienden significativamente. Este fenómeno puede explicarse por el hecho de que estos países presentan bajas temperaturas lo cual, posiblemente, fomente el desempeño de eventos de temporada de invierno y, por contra parte, sean menos competitivos en los de verano. No obstante, si bien Rusia y Alemania no tienen un medallero olímpico mayor al de EE.UU., sí mantienen una consistencia en su desempeño entre temporadas de invierno y verano.

III-B3. Atletas más influyentes: Con esta consulta se busca atender al rendimiento de aquellos atletas que resultaron determinantes e influyentes en la competitividad de sus países. Básicamente, la idea radica en establecer para cada país cuál es el atleta más influyente. Esta consulta, a su vez, entraña la dificultad de definir un criterio de medición de influencia para un atleta que contemple distintas variables y aspectos. Con este fin, se describen dos posibles criterios empleados.

Criterio 1: se considera a todos los atletas de un mismo país y se determina como más influyente aquel que haya obtenido más medallas de oro sobre la cantidad de veces que haya competido. En listado 5 se visualiza la consulta.

Listado 5 Consulta atletas más influyentes según medallas de oro ganadas.

```
MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED]->()
WITH c,a,COUNT(p) AS totalPerformances, COUNT(CASE p.medal WHEN 'Gold' THEN 1 END) AS goldMedals
ORDER BY c.name, goldMedals / toFloat(totalPerformances) DESC, totalPerformances DESC
WITH c, COLLECT({athlete: a.name, percentage: goldMedals / toFloat(totalPerformances),
totalPerformances: totalPerformances})[0] AS topAthlete
RETURN c.name AS Country, topAthlete.athlete AS Athlete,
topAthlete.percentage AS Percentage, topAthlete.totalPerformances AS Performances
ORDER BY Percentage DESC, Performances DESC
```

En la tabla II se muestran los atletas más influyentes para diez países. No obstante, este acercamiento puede considerarse un tanto *naive* dado que únicamente se tiene en cuenta las medallas de oro ganadas y no las de plata y bronce. Por ejemplo, para EE.UU. el atleta más influyente es Raymond Clarence con un total de diez medallas de oro sobre la diez veces que compitó. Sin duda representa números excelentes y demuestra una gran efectividad por parte del atleta. Sin embargo, en la tabla I se observa que el atleta Michael Phelps ha ganado un total de 23 medallas de oro sobre las 30 que compitió. Si bien la proporción de Clarence es mayor a la de Phelps, no se está considerando el hecho de que este último ganó, adicionalmente, tres medallas de plata y dos de bronce sumando un total de 28 medallas ganadas. De hecho, cualquier atleta que compita una vez y gane una medalla de oro ya posee la misma proporción que un atleta que haya competido seis veces y haya ganado cuatro de oro.

Por tanto, este criterio parece no ser muy asertivo ni tiene en cuenta las medallas de plata y bronce. Asimismo, se debería ponderar la cantidad de veces que haya competido un atleta, puesto que si este último logra un buen rendimiento en su disciplina, es probable que sea convocado para las siguientes competiciones, hecho que representa un factor muy meritorio.

Criterio 2: para este segundo criterio se define la siguiente función de ponderación.

Influencia Atleta = (Promedio Performance * 0.8) + (Competencias Ganadas), donde

Promedio Performance *= (Cantidad de medallas de oro + Cantidad de medallas de plata * 0.6 + Cantidad de medallas de bronce * 0.4) / Total Competencias*

Competencias Ganadas *= Cantidad de medallas de oro * 0.1 + Cantidad de medallas de plata * 0.05 + Cantidad de medallas de bronce * 0.03.*

El término *Promedio Performance* promedia el rendimiento de un atleta considerando las distintas medallas ganadas brindando una mayor ponderación a aquellas que sean de oro, mientras que *Competencias ganadas* es un factor de ponderación que beneficia en mayor proporción a aquellos atletas que hayan competido más veces y ganado alguna medalla de oro. La consulta se dispone en el listado 6.

Listado 6 Consulta atletas más influyentes según función de ponderación.

```
MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED]->()
WITH a, c, COUNT(p) AS totalPerformances,
COUNT(CASE p.medal WHEN 'Gold' THEN 1 END) AS goldMedals,
COUNT(CASE p.medal WHEN 'Silver' THEN 1 END) AS silverMedals,
COUNT(CASE p.medal WHEN 'Bronze' THEN 1 END) AS bronzeMedals
WITH toFloat(goldMedals + silverMedals*0.6 + bronzeMedals*0.4)/totalPerformances AS medalWeights,
totalPerformances, goldMedals, silverMedals, bronzeMedals, a, c
WITH ((goldMedals*0.1 + silverMedals*0.05 + bronzeMedals*0.03) + medalWeights*0.8) AS score, a, c
ORDER BY score DESC
WITH COLLECT({athlete:a.name,score:score})[0] AS topAthlete, c
RETURN c.name AS Country, topAthlete.athlete AS Athlete, round(topAthlete.score,3) AS Score
ORDER BY Score DESC
```

En la tabla III se observan los resultados asociados, donde se puede apreciar un cambio drástico de los mismos en comparación a la consulta anterior, lo cual expone una evidente mejora en la definición del criterio empleado. Notar que países como por ejemplo EE.UU., Italia, Rusia y Alemania, ahora poseen un nuevo atleta más influyente y probablemente más acertado.

Nota: en el anexo de este documento, se dispone de la definición de un tercer criterio V-A que resultó ser un paso intermedio para llegar a la definición final del criterio 2. Por otra parte, se muestra una tabla con los resultados de los atletas más influyentes sin considerar los países que representan V-B.

III-B4. Eventos olímpicos más variados: El objetivo de esta consulta es estudiar la variabilidad de los países ganadores dada un evento olímpico. La idea es evaluar el comportamiento de los eventos en su transcurso de las distintas ediciones de los juegos, y determinar si existe algún patrón respecto a sus equipos ganadores. La consulta se expone en el listado 7.

Listado 7 Consulta eventos más variados.

```
MATCH (e:Event)<-[p:PERFORMED]-()-[:FROM]->(c:Country)
WHERE p.medal = 'Gold'
RETURN e.name AS EventName, size(COLLECT(DISTINCT(c.name))) AS NumberOfCountries
ORDER BY NumberOfCountries DESC, EventName
```

En la tabla IV pueden observarse los eventos ordenados según su variabilidad en cuanto a países ganadores se refiere, siendo el fútbol y maratón de atletismo masculino los que destacan. Posiblemente se deba a la gran competitividad existente en dichos eventos. Sin embargo, existen otros donde sus ganadores son más estables en el tiempo como por ejemplo los que se

Country	Athlete	Percentage	Performances
USA	Raymond Clarence Ray Ewry"	1.0	10
Italy	Nedo Nadi	1.0	6
Hungary	Rudolf Krpti	1.0	6
Germany	Kristin Otto	1.0	6
Russia	Svetlana Alekseyevna Romashina	1.0	5
China	Chen Ruolin	1.0	5
Canada	Caroline Ouellette	1.0	4
Japan	Kaori Icho	1.0	4
Russia	Viktor Alekseyevich Krovopuskov	1.0	4
Sweden	Henri Julius Reverony Saint Cyr	1.0	4

Tabla II: Atleta más influyente por país, según proporción de medallas de oro ganadas.

observan en la tabla V, comportamiento que puede deberse a la gran destreza de los atletas de esos países en dichos eventos que desempeñan. Es de público conocimiento que existen países que se destacan en ciertos ámbitos o disciplinas olímpicas obteniendo resultados altamente satisfactorios, y esta consulta confirma la perpetuación de dichos resultados en el tiempo.

III-C. Algoritmos sobre grafos

En esta sección se pretende realizar distintas ejecuciones de algoritmos sobre grafos haciendo uso de la librería *Graph Data Science* con el fin de detectar nodos importantes (algoritmos de centralidad) y comunidades, dentro del grafo. En particular, la detección se centrará en nodos de atletas importantes y comunidades de estos mismos.

Primeramente, se debe realizar una proyección del grafo para luego poder ejecutar los algoritmos sobre dicha proyección. Para ello, se utiliza el procedimiento `gds.graph.project` que permite crear un nuevo grafo a partir de un grafo existente, seleccionando nodos y relaciones específicas [5]. Para el caso de estudio, la proyección efectuada intenta mantener las mismas condiciones del grafo original, manteniendo los nodos y relaciones pero sin incluir propiedades de estos últimos. Adicionalmente, para enriquecer el estudio, se optó por quitarle el sentido a las relaciones y que sean bidireccionales por medio de la directiva `UNDIRECTED`. En el listado 8 se encuentra la consulta que proyecta el nuevo grafo `myGraph`.

Listado 8 Consulta proyección de nuevo grafo.

Country	Athlete	Score
USA	”Michael Fred Phelps, II”	3.193
Russia	Larysa Semenivna Latynina (Diriy-)	1.843
Finland	Paavo Johannes Nurmi	1.77
Germany	Birgit Fischer-Schmidt	1.64
Italy	Edoardo Mangiarotti	1.513
Belgium	Gerard Theodor Hubert Van Innis	1.472
Jamaica	Usain St. Leo Bolt	1.44
UK	Jason Francis Kenny	1.404
Hungary	Rudolf Krpti	1.4
Norway	Ole Einar Bjrndalen	1.35

Tabla III: Atleta más influyente por país, según función de ponderación.

EventName	NumberOfCountries
Football Men’s Football	20
Athletics Men’s Marathon	17
Boxing Men’s Welterweight	15
”Wrestling Men’s Heavyweight, Greco-Roman”	15
”Wrestling Men’s Lightweight, Freestyle”	15
Athletics Men’s 5,000 metres	14
Boxing Men’s Featherweight	14
Çycling Men’s Road Race, Individual”	14
Rowing Men’s Double Sculls	14
Weightlifting Men’s Featherweight	14
”Wrestling Men’s Featherweight, Greco-Roman”	14
Athletics Men’s 1,500 metres	13
Boxing Men’s Bantamweight	13
Boxing Men’s Flyweight	13
Boxing Men’s Lightweight	13

Tabla IV: Eventos con países ganadores más variados.

EventName	Countries
Table Tennis Men’s Singles	South Korea,Sweden,China
Swimming Men’s 4 x 100 metres Medley Relay	Australia,USA
Diving Women’s Synchronized Platform	China

Tabla V: Eventos con países ganadores menos variados.

```
    CALL gds.graph.project(
  'myGraph',
  ['Athlete', 'Country', 'Event', 'Sport'],
  {
    FROM: {
      type: 'FROM',
      orientation: 'UNDIRECTED'
    },
    BELONGS_TO: {
      type: 'BELONGS_TO',
      orientation: 'UNDIRECTED'
    },
    PERFORMED: {
      type: 'PERFORMED',
      orientation: 'UNDIRECTED'
    }
  }
)
```

Las consultas de algoritmos de centralidad comparten la misma estructura, variando únicamente el nombre del algoritmo. En el listado 9 se observa su estructura.

Listado 9 Estructura genérica de consultas de algoritmos de centralidad.

```
CALL gds.<nombreAlgoritmo>.stream('myGraph')
YIELD nodeId, score
WITH gds.util.asNode(nodeId) AS node, score
WHERE labels(node) = ["Athlete"]
MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED]->()
WHERE a.name = node.name
WITH c, score, node, count(p) AS Performances
RETURN node.name AS Athlete, c.name AS Country,Performances, score AS Score
ORDER BY Score DESC, Country
LIMIT 15
```

III-C1. Algoritmo Page Rank: El algoritmo de centralidad *Page Rank* mide la importancia de cada nodo dentro del grafo, basándose en el número de relaciones entrantes y la importancia de los correspondientes nodos origen [6].

En la tabla VI se muestran los resultados de la ejecución. Se puede observar que todos los atletas tienen como factor común haber competido numerosas veces, en definitiva varias relaciones del tipo PERFORMED, lo cual refleja la definición del algoritmo. El tiempo de ejecución de la consulta fue de 1.8 segundos.

Nota: existen atletas que compitieron con más de una nacionalidad en distintas ediciones de los juegos, por lo cual un mismo atleta puede aparecer asociado a dos países distintos. Este factor, se discute más adelante.

III-C2. Algoritmo Article Rank: Este algoritmo de centralidad es una variante del Page Rank que mide la influencia transitiva de los nodos. Page Rank asume que las relaciones originadas desde nodos con menor *low-degree* (pocos arcos entrantes/salientes) tienen una influencia mayor que aquellas originadas desde nodos con *high-degree* (muchos arcos entrantes/salientes). Por el contrario, Article Rank disminuye la influencia de nodos con *low-degree* por medio de la disminución de los puntajes enviados a sus nodos vecinos, en cada iteración [7].

Athlete	Country	Performances	Score
"Michael Fred Phelps, II"	USA	30	3.3680405697994216
Eric Otto Valdemar Lemming	Sweden	23	3.150377121488921
Merlene Joyce Ottey-Page	Jamaica	19	2.6152641107972916
Merlene Joyce Ottey-Page	Slovenia	19	2.6152641107972916
Franziska van Almsick	Germany	23	2.614112599276352
"Henry William Furse Bill Hoskyns"	UK	21	2.5193108754390026
Oksana Aleksandrovna Chusovitina	Germany	29	2.5120435542107136
Oksana Aleksandrovna Chusovitina	Russia	29	2.5120435542107136
Oksana Aleksandrovna Chusovitina	Uzbekistan	29	2.5120435542107136
Rajmond Debevec	Serbia	20	2.4585604620332666
Rajmond Debevec	Slovenia	20	2.4585604620332666
Ole Einar Bjrndalen	Norway	27	2.4242916340998235
Heikki Ilmari Savolainen	Finland	39	2.40855656392878
Martina Moravcov	Czech Republic	19	2.3617338146141607
Martina Moravcov	Slovakia	19	2.3617338146141607

Tabla VI: Resultados ejecución Page Rank.

Los resultados de la ejecución de este algoritmo se visualizan en la tabla VII. Observándolos con detenimiento, los atletas que se posicionan en los primeros lugares, son aquellos que posee un alto grado de participación en las olimpiadas (más de veinte veces). Si se observan los resultados de la tabla VI, se puede notar que si bien se mantienen algunos nombres de atletas, aparecer otros nuevos, en especial en las primeras posiciones de la tabla como lo son Heikki Ilmari Savolainen y Joseph Josy Stoffe, que destacan por un alto número de competiciones, mientras que el nadador Michael Phelps pasa a estar en el tercer puesto, según los puntajes asignados. Parecería, entonces, que Article Rank, como bien explicita su definición, le asigna mayor importancia a aquellos nodos con un *high-degree* provocando que atletas que hayan competido numerosas veces (varios arcos o relaciones del tipo PERFORMED) posean un valor de *score* más alto, mientras que aquellos que hayan competido menor cantidad de veces, es decir nodos con un *low-degree*, se les asigne un menor *score*. Por otra parte, ateniendo nuevamente a los resultados de Page Rank, los que se encuentran en las primeras posiciones no son particularmente aquellos que hayan competido más veces; el atleta Heikki Ilmari Savolainen, decimotercero en la tabla VI, se encuentra en la primera posición en VII, mientras que Phelps se pasa de estar primero a colocarse tercero, en la segunda tabla. Este comportamiento, probablemente se deba a que el algoritmo Page Rank tiene en cuenta, asimismo, la importancia de los nodos origen. Posiblemente, Phelps al representar a EE.UU., un país que convoca una gran cantidad de atletas olímpicos en distintas disciplinas, posea un *score* asociado mayor que otro atleta que haya competido más veces. De hecho, si se observan los resultados de las consultas *Países con más medallas de oro* (III-B2) o *Atletas más influyentes* (III-B3), se llega a apreciar que países como EE.UU., Alemania, Rusia, Reino Unido y Noruega, siempre están presentes dado que corresponden a países con alto grado de participación en los juegos, hecho que conlleva a brindarle mayor relevancia a sus atletas.

La consulta tomó un tiempo estimado de aproximadamente 2 segundos.

III-C3. Algoritmo Degree Centrality: El presente algoritmo se utiliza para encontrar nodos populares dentro de un grafo. Básicamente, mide el número de relaciones entrantes o salientes (o ambas) de un nodo, dependiendo de la orientación de la relación proyectada [8].

Los resultados correspondientes se encuentran en la tabla VIII. En este caso, los atletas con mayor número de competiciones son directamente los que poseen mayor *score* siendo Heikki Ilmari el que compitió más veces dentro del conjunto total de atletas. A diferencia de Page Rank y Article Rank, no se pondera la influencia de otros nodos por lo cual el resultado no presenta mayor análisis. Notar que aquellos atletas con más de una nacionalidad están antes que otros con mayor cantidad de *Performances*. Esto se debe a que el score corresponde a la suma de la relaciones del tipo PERFORMED y FROM.

La consulta requirió de un tiempo de ejecución de aproximadamente 5.4 segundos, bastante mayor a las de Page y Article Rank.

III-C4. Algoritmo K-Means Clustering: Perteneciente a la familia de detección de comunidades, K-Means Clustering, es un algoritmo de aprendizaje no supervisado utilizado para resolver problemas de *clustering* (agrupamiento). Consiste de un simple procedimiento de clasificación de un conjunto de datos en un número de *clusters*, definido por un parámetro k. La librería GDS de Neo4j dirige el *clutering* basado en un *array* de *floats* que toma como entrada, que sea propiedad de algún nodo [9].

El objetivo de aplicación de este algoritmo consiste en encontrar comunidades de nodos de atletas que compartan características físicas similares, en especial relacionados a las propiedades de *height* (altura) y *weight* (peso). Para ello, se crea una nueva propiedad, *pyshique*, que se muestra en el listado 10. Asimismo, se realiza una nueva proyección del grafo en donde

Athlete	**Country**	**Performances**	**Score**
Heikki Ilmari Savolainen	Finland	39	0.6366726399159623
”Joseph Josy Stoffel”	Luxembourg	38	0.6229139168291492
”Michael Fred Phelps, II”	USA	30	0.607975019371277
Oksana Aleksandrovna Chusovitina	Germany	29	0.5837483107141431
Oksana Aleksandrovna Chusovitina	Russia	29	0.5837483107141431
Oksana Aleksandrovna Chusovitina	Uzbekistan	29	0.5837483107141431
Andreas Wecker	Germany	32	0.5677865248085304
Andreas Wecker	Germany	32	0.5677865248085304
Takashi Ono	Japan	33	0.5521244004044976
Michel Mathiot	France	31	0.5392270661805708
Ole Einar Bjrndalen	Norway	27	0.5367685604475055
Eric Otto Valdemar Lemming	Sweden	23	0.5320938940910664
Yordan Yovchev Yovchev	Bulgaria	30	0.5273319659157253
Karl Tore William Thoresson	Sweden	30	0.526341985091294
Franziska van Almsick	Germany	23	0.5029188195940121

Tabla VII: Resultados ejecución Article Rank.

Athlete	Country	Performances	Score
Heikki Ilmari Savolainen	Finland	39	40.0
”Joseph Josy Stoffel”	Luxembourg	38	39.0
Andreas Wecker	Germany	32	34.0
Andreas Wecker	Germany	32	34.0
Takashi Ono	Japan	33	33.0
Michel Mathiot	France	31	32.0
Oksana Aleksandrovna Chusovitina	Germany	29	32.0
Oksana Aleksandrovna Chusovitina	Russia	29	32.0
Oksana Aleksandrovna Chusovitina	Uzbekistan	29	32.0
Yordan Yovchev Yovchev	Bulgaria	30	31.0
Karl Tore William Thoresson	Sweden	30	31.0
”Michael Fred Phelps, II”	USA	30	31.0
Ole Einar Bjrndalen	Norway	27	28.0
Fabian Hambchen	Germany	26	27.0
Siegfried Flle	Germany	24	26.0

Tabla VIII: Resultados ejecución Degree Centrality.

únicamente se incluyen los nodos de atletas (Athlete) y la nueva propiedad creada physique. La consulta se dsipone en el listado 11.

Listado 10 Creación de la propiedad physique del nodo Athlete.

```
    MATCH (n:Athlete)
    SET n.physique = [toFloat(n.height), toFloat(n.weight)]
```

Listado 11 Consulta de proyección de un grafo con nodos Athlete y propiedad physique.

```
    CALL gds.graph.project(
    'athletes',
    {
      Athlete: {
        properties: 'physique'
      }
    },
    '*'
)
```

La estructura de esta consulta varía respecto a las anteriores y se muestra en el listado 12.

Listado 12 Consulta de ejecución del algorito K-Means Clustering.

```
    CALL gds.beta.kmeans.stream('athletes', {
    nodeProperty: 'physique',
    k: 5,
    randomSeed: 42
    })
    YIELD nodeId, communityId
    WITH gds.util.asNode(nodeId) AS node, communityId
    ORDER BY communityId, node.name ASC
    WITH communityId, COLLECT(node.name)[0..10] AS names,
    COLLECT(node.height)[0..10] AS heights, COLLECT(node.weight)[0..10] AS weights
    RETURN communityId, names AS Athletes, heights AS Heights, weights AS Weights
```

Los resultados de la consulta se ven en la tabla IX, en los cuales se perciben cinco comunidades distintas según la altura y peso de los atletas. Se observa que las comunidades fueron determinadas en base a un rango de valores de esas dos propiedades, donde en la comunidad con identificador 0 se encuentran los atletas con mayor altura, luego le sigue la comunidad con Id 2, luego la 4 y por último la 3. Asimismo, se puede notar que los valores de los pesos condicen con las alturas correspondientes. Por otra parte, la comunidad con Id 1 comprende todos aquellos atletas que originalmente presentaban valores nulos para dichas propiedades y por ende se optó por asignarles un valor predeterminado. Esto se dilucida con mayor énfasis en II-B2. La consulta realizada tomó un tiempo aproximado de 1 segundo.

communityId	Athletes	Heights	Weights
0	A. Joshua Josh West, Aaron E. Pollock, Aaron Kenneth Myette, Aaron Russell, Aaron Younger, Abas Arslanagi, Abbas Jadidi, Abbas Samimi, Abdel Aziz Boukar Boukar Moussa, Abdelkader Al-Zrouri	207,200,195,206,193,189,185,203,204,197	105,107,95,93,100,97,99,115,97,100
1	A. Albert,A. Dubois,A. Lawry,A. M. Woods,A. R. Upton,A. Willcocks,Aage Ernst Larsen,Aage Hy Pedersen,Aage Ingvar Eriksen,Aage Jrgen Christian Andersen	-1,-1,-1,-1,-1,-1,-1,-1,-1,-1	-1,-1,-1,-1,-1,-1,-1,-1,-1,-1
2	A Dijiang, Aagje Ada Kok (-van der Linden), Aapo Kustaa Perko, Aarik Wilson, Aarn Galindo Rubio, Aarne Ulf Kristian Lindroos, Aarne Vin Edward Honkavaara, Aaron Arthur Cook, Aaron Brown, Aaron Cleare	180,183,180,191,182,192,184,183,198,185	80,85,82,88,80,88,81,80,79,84
3	Th Anh, Th Ngn Thng, Tin Tun, A Lamusi, A. W. Nancy Nan Rae, Aagje Vanwalleghem, Aarne Eino Henrik Vehkonen, Aarne Kainlauri, Aaron Benjamin Herman, Aaron Dupnai	165,147,173,170,156,155,160,171,153,162	58,47,63,60,53,43,57,63,50,54
4	A. Abdul Razzak, A. J. Tyronne Benildus Benny Fernando, Aadam Ismaeel Khamis, Aadolf Fredrik Svanstrm, Aafke Hament, Aage Birch, Aage Brge Poulsen, Aage Carl Christian Lassen, Aage Rasmussen (-Remfeldt), Aaltje Grietje Alie Boorsma	178,179,172,179,181,172,185,181,177,179	70,70,67,70,64,70,68,62,67,75

Tabla IX: Resultados ejecución K-Means Clustering.

IV. Conclusiones y trabajo futuro

Como compendio de lo elaborado en el presente documento, se diseño e implementó una base de datos de grafos en base al dominio de los Juegos Olímpicos mediante la utilización de la herramienta Neo4j, efectuando uso de sus capacidades y funcionalidades que esta provee, como son su potente lenguaje de consulta *Cypher* y su librería *Graph Data Science* que ofrece una amplia gama de algoritmos sobre grafos. El proceso de desarrollo llevado a cabo implicó tareas como la obtención de los datos, su estudio y análisis, un pre procesamiento de los mismos con el fin de filtrar y mantener aquellos datos que aporten mayor valor semántico al propósito de estudio. Consiguientemente, se prosiguió con la carga de dichos datos a Neo4j desde la plataforma misma, lo cual, en primera instancia, no se tuvo éxito debido al gran volumen de datos tratado. De modo de solventar esta problemática, se realizó la carga utilizando el lenguaje Python junto con la librería *neo4j* para la conexión a la base. Esta alternativa, resultó ser exitosa constatando la flexibilidad y usabilidad de la librería para conectarse a una base en Neo4j. Asimismo, los objetivos especificados en la sección de Introducción I fueron satisfechos con éxitos.

IV-A. Análisis de los datos

A través de las consultas efectuadas a lo largo de este documento, se relevaron estadísticas de interés en relación a los Juegos Olímpicos, evaluando resultados, relacionándolos y estudiando el por qué de ciertos fenómenos o acontecimientos vinculados a los datos. Particularmente, consultas como las de *Países con más medallas de oro* o *Atletas más influyentes* permitieron plasmar un mejor entendimiento sobre el desempeño y competitividad de ciertos países a lo largo de la historia de los juegos, así como también comprender el hecho de por qué ciertos países siempre tienden a permanecer en la cima en ciertas disciplinas. Se concluye que EE.UU., Rusia, Alemania, Reino Unido son principalmente los países con mayor influencia en los Juegos Olímpicos, existiendo otros como China que han incrementado su competitividad, involucrándose en distintas disciplinas o eventos olímpicos.

Asimismo, se evaluó la enorme capacidad y flexibilidad del lenguaje de consultas *Cypher* que brinda una gran versatilidad para el modelado de consultas complejas así como también un gran abanico de funcionalidades.

IV-B. Librería Graph Data Science

La ejecución de algoritmos sobre grafos conformó un componente esencial en el desarrollo de este estudio. Particularmente, se estudió la detección de nodos importantes y comunidades. Para el primero, se ejecutaron tres algoritmos que arrojaron resultados semánticamente distintos según el mecanismo de asignación de puntajes que implementa cada algoritmo, siendo Page Rank el que toma en mayor consideración la importancia de cada nodo. Por otra parte, se ejecutó el algoritmo K-Means Clustering para la detección de comunidades de atletas según las propiedades de peso y altura de estos. Se concluye el gran poder de discernimiento del algoritmo.

Asimismo, el tiempo de ejecución de estos algoritmos es realmente estremecedor en relación al gran tamaño de los datos con los que se trabaja, siendo Degree Centrality el que requirió de mayor tiempo de ejecución (5.4 segundos). Esto demuestra la eficiente implementación de estos algoritmos por parte de GDS sobre grandes cantidades de datos, optimizando tanto los tiempos de ejecución como el uso de memoria.

IV-C. Rendimiento de Neo4j

Habiéndose efectuado análisis y consultas de distintos tipos así como ejecución de diversos algoritmos, es factible concluir la gran capacidad y rendimiento que posee Neo4j para la implementación de base de datos de grafos y su facilidad para el manejo y soporte de grandes volúmenes de datos. En este sentido, se destaca además el modelado de consultas complejas a través de su lenguaje *Cypher* y la ejecución de algoritmos sobre grafos por medio de la librería GDS.

IV-D. Trabajo a futuro

De modo de continuar con los lineamientos del proyecto, se plantea la posibilidad de un mejor manejo de los datos para el caso de atletas con más de una nacionalidad. Existen atletas que están asociados a más de un país dentro del conjunto de datos debido a que los mismos presentan una doble o hasta triple nacionalidad, pudiendo haber competido en distintas ediciones, para más de un país. Asimismo, existen otros casos de atletas donde aparecen asociados más de una vez al mismo país. Este extraño fenómeno se debe a que históricamente algunos países, debido a circunstancias de guerra y política, estaban divididos internamente, es decir si bien el nombre del país es el mismo, el NOC puede variar según sea la división. Por ejemplo, Alemania es representado mediante el NOC ”GER” mientras que Alemania del Este, ”GDR” y Alemania Oeste, ”FRG”. Todos figuran como Alemania pero su NOC varía. Por tanto, se cree que debería definirse un criterio justo para una fiel y coherente representación de los datos respecto de la realidad. A modo de ejemplo, podría representarse Alemania con el NOC ”GER” y englobar a todos los demás NOC dentro de uno solo.

IV-E. Conclusión final

El trabajo realizado permitió un primer acercamiento al dominio de la bases de datos de grafo, así como comprender su gran flexibilidad para el manejo y soporte de grandes cantidades de datos. El hecho de no tener un esquema asociado, como sucede en el caso de las bases relaciones, posibilita la libre elección de estructuración de los datos según sea lo que se pretende modelar o las consultas que se quieren responder, como lo fue en este caso.

V. Anexo

V-A. Un tercer criterio de influencia

Evaluando qué criterio sería el más adecuado, en un transcurso intermedio al criterio final expuesto en el documento, se definió el siguiente. Se considera para cada país cuál fue el atleta que compitió más veces y el que menos compitió. Luego para cada atleta de un país dado se efectúa el siguiente cálculo.

Influencia Atleta = (Cantidad de medallas de oro * 3 + Cantidad de medallas de plata * 2 + Cantidad de medallas de bronce) * Ponderación Competencias

,donde

Ponderación Competencias = *(Total Competencias - Atleta País Con Menos Competencias) / (Atleta País Con Más Competencias - Atleta País Con Menos Competencias).*

En el listado 13 se visualiza la consulta realizada.

Listado 13 Consulta tercer criterio de atleta más influyente.

```
1     MATCH (c:Country)<-[:FROM]-(a:Athlete)-[p:PERFORMED]->()
2     WITH c.name AS country, a, COUNT(p) AS cant
3     WITH country, MIN(cant) AS minPerformances, MAX(cant) AS maxPerformances
4     ORDER BY country
5     WITH COLLECT({country:country, minPerformances:minPerformances, maxPerformances:maxPerformances}) AS data
6
```

```
7
8       MATCH (c2:Country)<-[:FROM]-(a2:Athlete)-[p2:PERFORMED]->()
9       WITH c2, a2, COUNT(p2) AS TotalPerformances,
10          COUNT(CASE p2.medal WHEN 'Gold' THEN 1 END) AS GoldMedals,
11          COUNT(CASE p2.medal WHEN 'Silver' THEN 1 END) AS SilverMedals,
12          COUNT(CASE p2.medal WHEN 'Bronze' THEN 1 END) AS BronzeMedals,
13          [countryData IN data WHERE countryData.country = c2.name | countryData.minPerformances][0] AS MinPerformances,
14          [countryData IN data WHERE countryData.country = c2.name | countryData.maxPerformances][0] AS MaxPerformances
15      WITH c2, a2, TotalPerformances,
16          GoldMedals, SilverMedals, BronzeMedals,
17          MinPerformances, MaxPerformances,
18          CASE WHEN (MaxPerformances - MinPerformances) = 0
19              THEN toFloat((GoldMedals * 3 + SilverMedals * 2 + BronzeMedals)) *
20                  (TotalPerformances - MinPerformances) / 1
21              ELSE toFloat((GoldMedals * 3 + SilverMedals * 2 + BronzeMedals)) *
22                  (TotalPerformances - MinPerformances) / (MaxPerformances - MinPerformances)
23          END AS Weight
24      ORDER BY c2.name, Weight DESC
25      WITH c2.name AS Country, COLLECT({athlete: a2.name, weight: Weight})[0] AS TopAthlete
26      RETURN Country, TopAthlete.athlete AS Athlete, round(TopAthlete.weight,3) AS Weight
27      ORDER BY Weight DESC
```

Los resultados se observan en la tabla X.

V-B. Atletas más influyentes

Se muestra una tabla con los veinte atletas más influyentes según el criterio 2 definido en el documento pero sin tener en cuenta los países sino únicamente los nodos atletas y sus desempeños. En la tabla XI se visualizan los resultados junto a las estadísticas de cada atleta.

Country	Athlete	Weight
USA	”Michael Fred Phelps, II”	77.0
Norway	Ole Einar Bjrndalen	33.0
Japan	Takashi Ono	27.0
Belgium	Gerard Theodor Hubert Van Innis	18.0
Belarus	Vitaly Venediktovich Shcherbo	17.368
Netherlands	”Theodora Elisabeth Gerarda Anky van Grunsven”	16.25
Italy	Edoardo Mangiarotti	15.0
Zimbabwe	Kirsty Leigh Coventry (-Seward)	15.0
Hungary	Aladr Gerevich (-Gerei)	14.13
Finland	Heikki Ilmari Savolainen	14.0

Tabla X: Resultados definición de un tercer criterio de influencia.

Athlete	Country	score	goldMedals	silverMedals	bronzeMedals	totalPerformances
”Michael Fred Phelps, II”	USA	3.193	23	3	2	30
Larysa Semenivna Latynina (Diriy-)	Russia	1.843	9	5	4	19
Raymond Clarence Ray Ewry”	USA	1.8	10	0	0	10
Paavo Johannes Nurmi	Finland	1.77	9	3	0	12
”Frederick Carlton Carl Lewis”	USA	1.718	9	1	0	10
Mark Andrew Spitz	USA	1.647	9	1	1	12
Birgit Fischer-Schmidt	Germany	1.64	8	4	0	13
Birgit Fischer-Schmidt	Germany	1.64	8	4	0	13
”Matthew Nicholas Matt Biondi”	USA	1.57	8	2	1	12
Viktor Ivanovych Chukarin	Russia	1.549	7	3	1	11
Edoardo Mangiarotti	Italy	1.513	6	5	2	13
”Donald Arthur Don Schollander”	USA	1.51	7	1	0	8
Gerard Theodor Hubert Van Innis	Belgium	1.472	6	4	0	10
Isabelle Regina Werth	Germany	1.472	6	4	0	10
”Jennifer Elisabeth Jenny Thompson (-Cumpelik)”	USA	1.46	8	3	1	17
Usain St. Leo Bolt	Jamaica	1.44	8	0	0	10
Nikolay Yefimovich Andrianov	Russia	1.413	7	5	3	24
Jason Francis Kenny	UK	1.404	6	1	0	7
Nedo Nadi	Italy	1.4	6	0	0	6
Rudolf Krpti	Hungary	1.4	6	0	0	6

Tabla XI: Resultados atletas más influyentes.

REFERENCIAS

[1] Ruben Cañadas. Base de datos orientada a grafos o graph database. Sitio web, 01 2022. https://abdatum.com/informatica/base-datos-orientada-grafos.

[2] Lorena Etcheverry. Diseño y modelado en bases de datos de grafos. Sitio web, 2022. https://eva.fing.edu.uy/mod/resource/view.php?id=136475.

[3] rgriffin. 120 years of olympic history: athletes and results. Sitio web, Mayo 2018. https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results?select=athlete_events.csv.

[4] Santiago Costa. Bdnr olympic stats, Junio 2023. https://gitlab.fing.edu.uy/santiago.costa/bdnr-olympic-stats.

[5] Projecting graphs using native projections. https://neo4j.com/docs/graph-data-science/current/management-ops/projections/graph-project/.

[6] Pagerank. https://neo4j.com/docs/graph-data-science/current/algorithms/page-rank/.

[7] Article rank. https://neo4j.com/docs/graph-data-science/current/algorithms/article-rank/.

[8] Degree centrality. https://neo4j.com/docs/graph-data-science/current/algorithms/degree-centrality/.

[9] K-means clustering. https://neo4j.com/docs/graph-data-science/current/algorithms/kmeans/.