

Rapport de Projet



NEO4J : Modélisation d'une BDD

Mamady DJIGUINE

Nouh CHELGHAM

Université Toulouse III Paul Sabatier

MASTER 2 IAFA, SEMESTRE 9

Prof : PINEL-SAUVAGNAT Karen

Année 2024/2025

1 Schéma

Dans le cadre de ce projet en Neo4j, l'objectif est de concevoir et de modéliser une base de données orientée graphe qui représente un dataset détaillé des Jeux Olympiques. Pour commencer, une étape cruciale consiste à élaborer un schéma de données adapté, afin de structurer efficacement les relations entre les différentes entités, et permettre une exploration fluide des données. En utilisant Neo4j, nous pouvons modéliser des éléments clés tels que les athlètes, les pays, les sports, les éditions des Jeux, les médailles remportées, et les tweets relatifs aux compétitions. Chaque édition des Jeux devient un nœud central, lié aux pays participants, aux athlètes et aux sports disputés. Les médailles sont associées à des athlètes et à des pays, ce qui permet de visualiser les performances et les succès au fil des années. De plus, les tweets peuvent être intégrés comme des nœuds associés aux événements et athlètes, ajoutant une dimension dynamique aux Jeux Olympiques. Ce schéma de données nous offrira une base solide pour explorer les relations et dynamiques complexes au sein du dataset, et pour interroger efficacement les données en fonction des objectifs du projet.

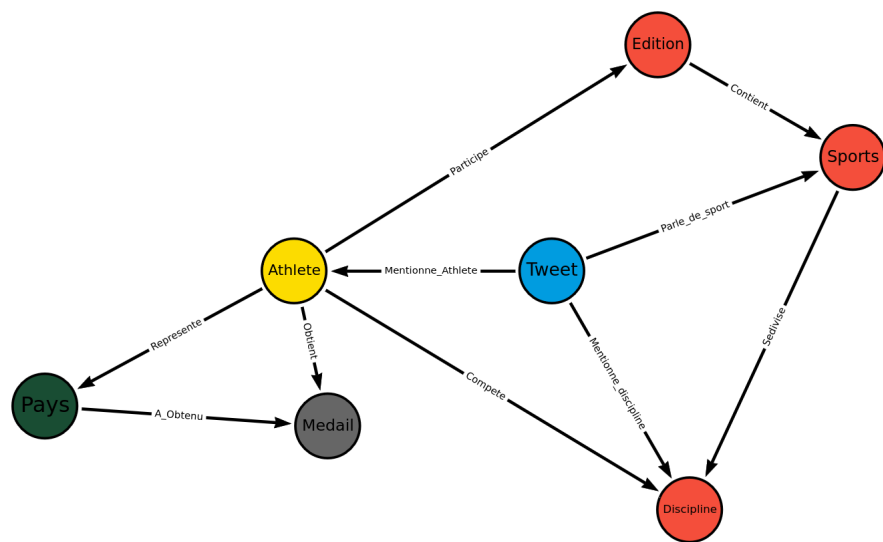


Figure 1: Schéma de Modélisation

Les nœuds et les relations définis dans ce projet permettent de représenter de manière claire et structurée toutes les entités et leurs interactions. Par exemple, un ATHLÈTE participe à une édition, et chaque édition est composée de sports, lesquels sont eux-mêmes divisés en disciplines. Le nœud MEDAL, avec trois instances distinctes (Gold, Silver, Bronze), standardise les types de médailles, facilitant ainsi les requêtes sur les médailles gagnées.

Les TWEETS sont associés aux athlètes, aux sports ou aux disciplines via

des hashtags. Cela permet de filtrer les tweets pertinents pour chaque entité et d'effectuer des requêtes afin d'identifier, par exemple, les athlètes les plus mentionnés pendant une édition des Jeux.

Les nœuds PAYS permettent de regrouper facilement les données par pays, ce qui est essentiel pour les analyses sur les pays sans nécessiter de recalculer le tableau des médailles à chaque requête. Pour mieux gérer, par exemple, le nombre de médailles gagnées par un pays et leur répartition par type, une relation a été ajoutée entre les pays et les médailles, avec des propriétés conservant des informations sur la discipline, l'édition, et le nombre total de chaque type de médaille.

Afin de gérer le pays de naissance des athlètes indépendamment des pays qu'ils représentent, le pays de naissance est conservé comme une propriété du nœud ATHLETE. Une relation supplémentaire, Représente, est également ajoutée entre les pays et les athlètes pour gérer les pays que chaque athlète représente.

Enfin, pour mieux suivre les résultats des athlètes et éviter des requêtes complexes, une relation Obtient est établie entre le nœud ATHLETE et MEDAL, avec des propriétés permettant de gérer l'historique des résultats des athlètes dans les épreuves des éditions auxquelles ils ont participé.

2 Importation

pour l'étape de l'importation, un notebook avec des commentaires et des guides sera déposé sur moodle pour montrer les étapes faites pour le nettoyage des données.

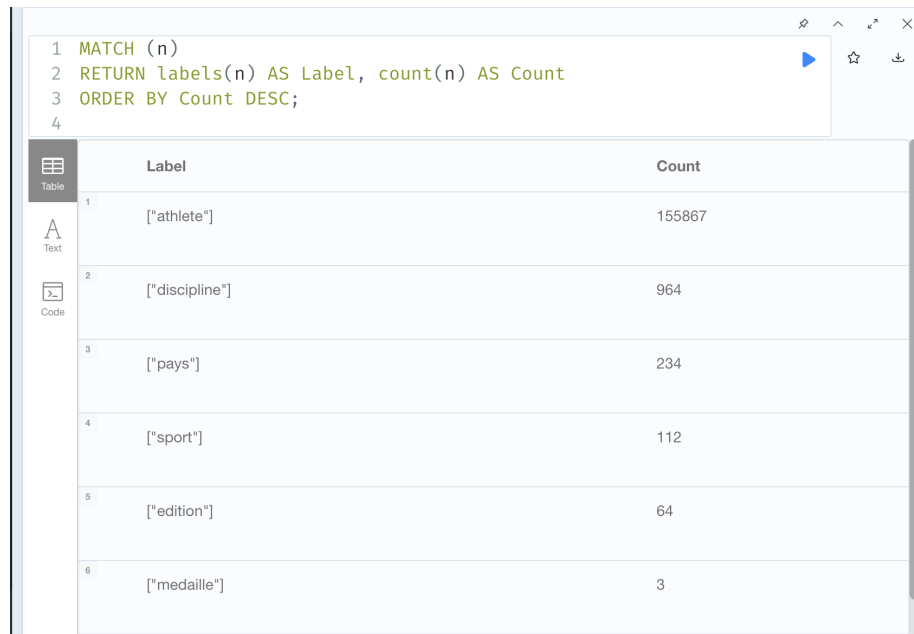
```
..... 90% Δ73ms [1s 605ms]
..... 95% Δ47ms [1s 653ms]
..... 100% Δ45ms [1s 698ms]
Imported 1 068 553 relationships in 3s 736ms
Flushing stores
Flush completed in 201ms
IMPORT DONE in 8s 20ms.
(base) djiginemamady@MacBook-Pro-de-DJIGUINE dbms-4041231d-0d44-4e36-86aa-e8ade8c707
d9 %
```

```
..... 90% Δ0ms [3s 198ms]
..... 95% Δ0ms [3s 198ms]
..... 100% Δ0ms [3s 198ms]
Imported 164 840 nodes in 3s 201ms
Prepare ID mapper
..... 5% Δ47ms [47ms]
..... 10% Δ45ms [51ms]
```

3 Requetes Cypher

Pour les requetes demandées, voici des capture d'écrans avec les résultats de ces requetes sur notre base de données

3.1 le nombre de nœuds par label



The screenshot shows a Cypher query interface. The query is as follows:

```
1 MATCH (n)
2 RETURN labels(n) AS Label, count(n) AS Count
3 ORDER BY Count DESC;
4
```

The results are displayed in a table with two columns: 'Label' and 'Count'. The table is ordered by 'Count' in descending order.

	Label	Count
1	["athlete"]	155867
2	["discipline"]	964
3	["pays"]	234
4	["sport"]	112
5	["edition"]	64
6	["medaille"]	3

3.2 le nombre de relations par type

```
1 MATCH ()-[r]→()
2 RETURN type(r) AS RelationType, count(r) AS Count
3 ORDER BY Count DESC;
```

RelationType	Count
"Compete"	241095
"Participe"	215804
"Represente"	158126
"Obtient"	44687
"A_obtenu"	20384
"Contient"	1138
"Sedivise"	1094

3.3 les athlètes qui ont gagné une médaille à l'épreuve Decathlon, Men en 2020

```
1 MATCH (a:athlete)-[r:Obtient]→(m:medaille)
2 MATCH (e:edition {edition_id: r.edition_id})
3 WHERE e.year = '2020' AND r.discipline = "Decathlon, Men"
4 RETURN a.name AS athlete, a.country AS Pays, m.medal AS
   Médaille, e.edition AS
   Edition, e. year AS Annee;
```

	athlete	Pays	Medaille	Edition	Annee
1	"Damian Warner"	"Canada"	"Gold"	"2020 Summer Olympics"	"2020"
2	"Kévin Mayer"	"France"	"Silver"	"2020 Summer Olympics"	"2020"
3	"Ashley Moloney"	"Australia"	"Bronze"	"2020 Summer Olympics"	"2020"

3.4 le nombre d'athlètes féminines en 2016

```
1 MATCH (a:athlete)-[:Participe]→(e:edition {year: '2016'})
2 WHERE a.sex = "Female"
3 RETURN count(a) AS NombreAthletesFeminines;
```

	NombreAthletesFeminines
1	5137

3.5 les athlètes qui ont participé aux jeux pour un pays dans lequel ils ne sont pas nés

```
1 MATCH (a:athlete)-[:Represente]→(p:pays)
2 WHERE a.country_noc <> p.country_noc
3 RETURN a.name AS Athlete, a.country AS PaysNaissance,
p.country AS PaysRepresente LIMIT(10);
```

Athlete	PaysNaissance	PaysRepresente
"Antonio Oliveira"	"Brazil"	"Argentina"
"Frank Beaurepaire"	"Australasia Australia"	"Australia"
"Sergo Chakhoyan"	"Armenia Australia"	"Australia"
"Aleksan Karapetyan"	"Armenia Australia"	"Australia"
"Sahit Prizreni"	"Albania Australia"	"Australia"
"Charles Granville Bruce"	"Great Britain"	"Australia"
"Bill Strutt"	"Great Britain"	"Australia"
"George Mallory"	"Great Britain"	"Australia"
"Henry Morshead"	"Great Britain"	"Australia"
"John Noel"	"Great Britain"	"Australia"

3.6 les tweets de l'édition 2020 qui concernent le nageur Michael Phelps (hashtag michaelphelps)

```
1 MATCH (t:tweet)-[:T_mentions_Athlete]→(a:athlete)
2 WHERE ANY(hashtag IN t.hashtags WHERE toLower(hashtag)
   CONTAINS '2020')
3 AND ANY(hashtag IN t.hashtags WHERE toLower(hashtag) CONTAINS
   'michaelphelps')
4 RETURN t.id AS idTweet, t.hashtags AS Hashtags, a.name AS
   NomAthlete
5
```

idTweet	Hashtags	Nom
"t1419878616511430656"	"['michaelphelps', 'tokyo2020']"	"Mi
"t1419150908793884673"	"['tokyo2020', 'olympics', 'ussainbolt', 'michaelphelps']"	"Mi
"t1419838116228173824"	"['michaelphelps', 'tokyo2020']"	"Mi
"t1419857652071419912"	"['michaelphelps', 'olympics', 'tokyo2020']"	"Mi
"t1419192997841969152"	"['tokyo2020', 'olympics', 'michaelphelps', 'olympicgames']"	"Mi

3.7 les disciplines et les sports associés qui ont été proposées sur moins de 10 éditions

```
1 MATCH (d:discipline)←[:Sedivise]-(s:sport)←[:Contient]-(e:edition)
2 WITH d,s, count(DISTINCT e) AS NombreEditions
3 WHERE NombreEditions < 10
4 RETURN d.discipline AS Discipline, s.sport AS Sport,
NombreEditions LIMIT(10)
```

Discipline	Sport	NombreEditions
"Tug-Of-War, Men"	"Tug-Of-War"	5
"Individual, Women"	"Golf"	4
"Individual, Men"	"Golf"	4
"Team, Men"	"Golf"	4
"Individual, Handicap, Men"	"Golf"	4
"Driving Contest, Men"	"Golf"	4
"Individual, Professional, Men"	"Golf"	4
"Putting Contest, Men"	"Golf"	4
"Frontenis, Doubles, Men"	"Basque pelota"	4
"Cesta Punta, Doubles, Men"	"Basque pelota"	4

4 DashBoard

Pour la dernière étape de ce projet, un fichier .json sera déposé sur Moodle. Ce fichier proposera une visualisation des informations clés identifiées, telles que le pourcentage d'athlètes participant à plusieurs disciplines ou encore la répartition des médailles par pays. (une capture qui represente les visualisations clés, d'autres sont dans le fichier .json)

