PROJET Neo4J

1. Description des besoins

Le cas d'étude proposé concerne les jeux olympiques. Les données utilisées sont plus précisément décrites dans le document « Cas d'étude – Jeux Olympiques ».

Nous souhaitons construire un graphe de connaissance (à la « Google Knowledge Graph »[1]) concernant toutes les éditions des jeux olympiques, les athlètes participants, et les résultats obtenus. Ce graphe de connaissance doit pouvoir être étendu par la suite.

De manière très résumée, chaque édition des jeux olympiques est :

- organisée par une ville,
- composée de sports (ex : basket, athlétisme), eux-mêmes divisés en disciplines (exemple : basket féminin / basket masculin ; 100m/400m/décathlon/saut en hauteur). Les différentes disciplines peuvent avoir lieu sur plusieurs journées d'épreuves (event), chaque journée se déroulant sur un site.

Les athlètes, représentant un pays, obtiennent des résultats sur les différentes journées d'épreuves, ces résultats pouvant être associés à des médailles. Depuis les années 2010, des tweets sur les disciplines/sports/athlètes, viennent compléter et illustrer les différentes performances.

À l'aide du graphe créé, il s'agira par exemple d'être capable d'avoir les informations suivantes :

- Pour un athlète : donner sa biographie, ses médailles, les éditions auxquelles il a participé, ses plus grands concurrents, les tweets qui l'ont mentionné, …
- Pour un pays : de lister les sports et disciplines dans lesquels il a obtenu le plus de médailles, …
- Pour un sport ou une discipline : de connaître les meilleurs sportifs dans le temps, les tweets publiés sur les dernières éditions,
- Pour une édition : d'avoir les chiffres significatifs (médailles, épreuves, athlètes qui ont marqué l'édition, etc.)

L'idée est d'être capable de produire des « fiches résumées » sur des éditions, sports, disciplines, pays ou athlètes. Un exemple de ce qui est fait avec le Google Knowledge Graph est présenté en Annexe pour l'athlète Marie-José Pérec.

On souhaite également pouvoir donner quelques statistiques d'évolution au cours du temps, pour les éditions en général ou pour une discipline / un sport spécifique (par exemple évolution du nombre d'épreuves au cours des éditions d'été, évolution du nombre d'athlètes féminines participant, etc).

[1] https://en.wikipedia.org/wiki/Google_Knowledge_Graph

2. Modélisation

À faire : Proposez une modélisation associée au cas d'étude, **en justifiant vos choix.**

Les contraintes suivantes sont imposées dans votre réflexion :

- On ne garde que les tweets qui ont des hashtags correspondant à un athlète, sport ou discipline
- Il doit y avoir un nœud de label MEDAILLE avec 3 instances (Gold, Silver, Bronze)
- Il doit y avoir des nœuds de label SPORT et DISCIPLINE
- Comme de nombreuses analyses vont être faites par pays, il n'est pas souhaitable de recalculer le tableau des médailles à chaque requête

3. Importation des données

À faire : Créez un notebook python permettant de transformer les données en entrée sous la forme nécessaire à l'importation Neo4j.

Il faut un fichier .csv par label de nœud et un fichier .csv par type de relation.

Exemples :
Nœuds city

city.csv

```
city_id;city_name
ci1,Athina
ci2,Paris
ci3,St. Louis
```

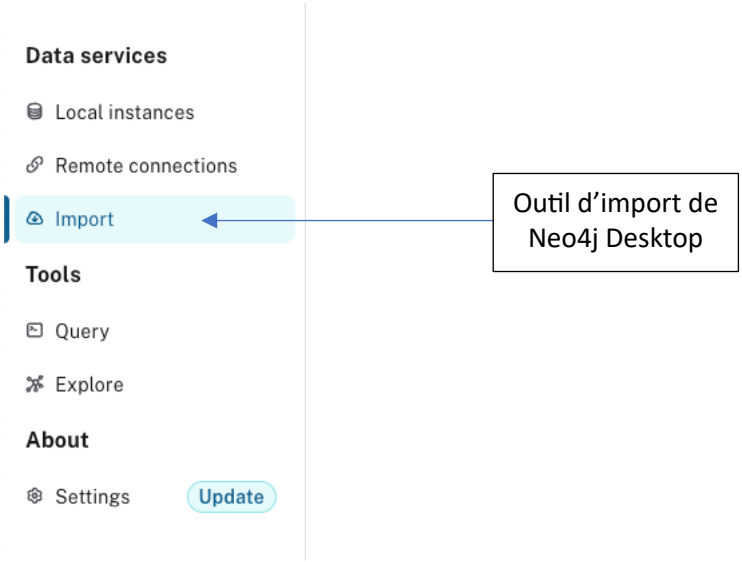
Relation in_country

in_country.csv

```
city_id;country_id
ci1,GRE
ci2,FRA
ci3,USA
```

À faire : Importer les données précédemment créées sous Neo4j.

Pour importer les données, vous allez utiliser l'outil d'import de Neo4J Desktop.



Vous trouverez la documentation de l'import dans l'archive `exemple.zip` disponible sous Moodle, qui contient également un petit exemple de données à importer.

Remarques :

- il est fortement conseillé de s'entrainer sur l'exemple (avec tous les fichiers .csv présents dans le répertoire '`pour chargement/via outil import`') avant de basculer sur une grande quantité de données.
- Avant de commencer l'import de vos données, n'oubliez pas de créer une nouvelle base sur votre instance.

4. Requêtes Cypher

À faire : Formuler et exécuter les requêtes Cypher suivantes sur votre base.

1. Donner le nombre de nœuds par label ;
2. Donner le nombre de relations par type ;
3. Donner les athlètes (nom, pays représenté) qui ont gagné une médaille à l'épreuve « Decathlon, Men » en 2020 ;
4. Donner le nombre d'athlètes féminines en 2016 ;
5. Donner tous les athlètes qui ont participé aux jeux pour un pays dans lequel ils ne sont pas nés ;
6. Donner les tweets de l'édition 2020 qui concernent le nageur Michael Phelps (hashtag michaelphelps) ;
7. Donner les disciplines (et les sports associés) qui ont été proposées sur moins de 10 éditions.

5. Visualisation des données

À faire : Programme Python permettant l'interrogation du graphe et la visualisation des données

Nous souhaitons faire un programme d'affichage et de visualisation des données, qui servira de *proof-of-concept* pour les fiches descriptives des éditions/sports/athlètes/disciplines qui devront être faites à terme[2].

Pour ce faire, vous allez utiliser Python et 2 bibliothèques spécifiques :

- `neo4j` pour l'interaction avec le graphe : envoi de requêtes et récupération de résultats
- une bibliothèque de visualisation (à votre convenance, comme par exemple `matplotlib`, `pyplot` ou `seaborn`)

L'archive `exemple.zip` contient dans le répertoire '`pour visualisation`' un exemple de code python utilisant le jeu de données du répertoire '`pour chargement/via outil import`'. Vous avez également les requêtes Cypher pour création directe du graphe d'exemple dans le répertoire '`pour chargement/via Cypher`'.

Les possibilités étant très larges, vous devrez choisir si vous souhaitez afficher des graphiques/infos pour :

- un athlète
- une discipline
- un sport
- une édition des jeux olympiques

À vous de choisir les requêtes/visualisations qui mettront le mieux vos données en valeur. Au moins 5 visualisations sont attendues pour l'athlète/sport/discipline/édition que vous avez choisi.
Idéalement, votre programme sera paramétrable et permettra à l'utilisateur d'entrer un nom d'athlète/discipline/sport/année d'édition avant d'avoir les visualisations.

6. À rendre pour le projet

Vous devrez rendre une archive .zip avec :

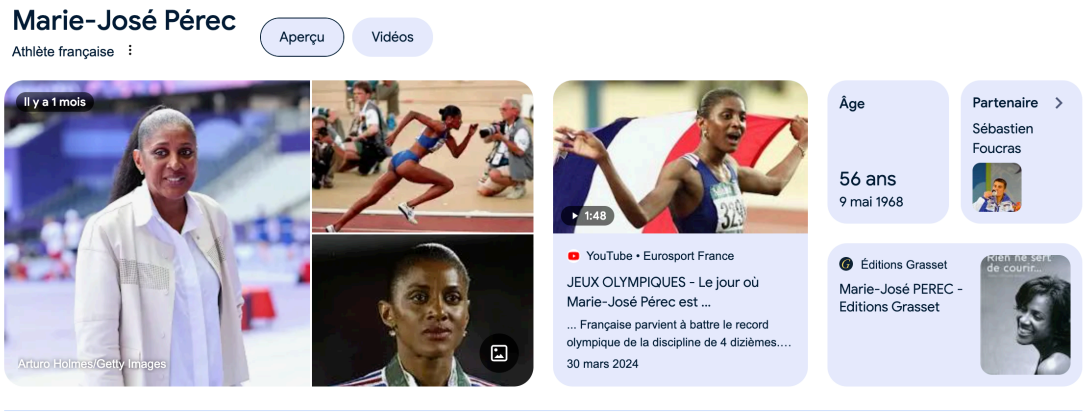
- Un rapport PDF contenant
  - La modélisation retenue et sa **justification**
  - Une visualisation du schéma de votre graphe[3] + d'un extrait des données
  - Les requêtes Cypher demandées (avec une capture d'écran du résultat)
  - Des captures d'écran des graphiques générés à l'étape de visualisation
- Votre notebook Python de transformation des données (Étape importation des données)
- Votre programme python de visualisation (Etape visualisation).

[2] Vous n'avez pas à faire ces fiches pour le projet.

[3] `CALL db.schema.visualization()`

7. Annexe

Exemple de « fiche » descriptive d'une athlète, obtenue ici à la suite d'une recherche Google basée sur le Google Knowledge Graph. Dans notre cas, la fiche n'utilisera que les informations existantes dans le jeu de données initial.



À propos

Marie-José Pérec, née le 9 mai 1968 à Basse-Terre en Guadeloupe, est une athlète française. Elle est la seule Française triple championne olympique d'athlétisme : en 1992 aux Jeux de Barcelone et deux fois aux Jeux d'Atlanta en 1996. Wikipédia

Date/Lieu de naissance : 9 mai 1968 (Âge: 56 ans), Basse-Terre, Guadeloupe

Partenaire : Sébastien Foucras

Médaille D'or : 400 mètres féminin aux Jeux olympiques d'été de 1992, PLUS

Parents : Josette Pérec

Taille : 1,8 m

Distinctions : ESPY Award de la meilleure athlète d'athlétisme

Club : Paris UC, Stade français

Commentaires

Profils

Instagram

X (Twitter)

Les internautes recherchent aussi

Tony Estanguet

Tendances

Sébastien Foucras

Teddy Riner

Thomas Jolly

Tendances