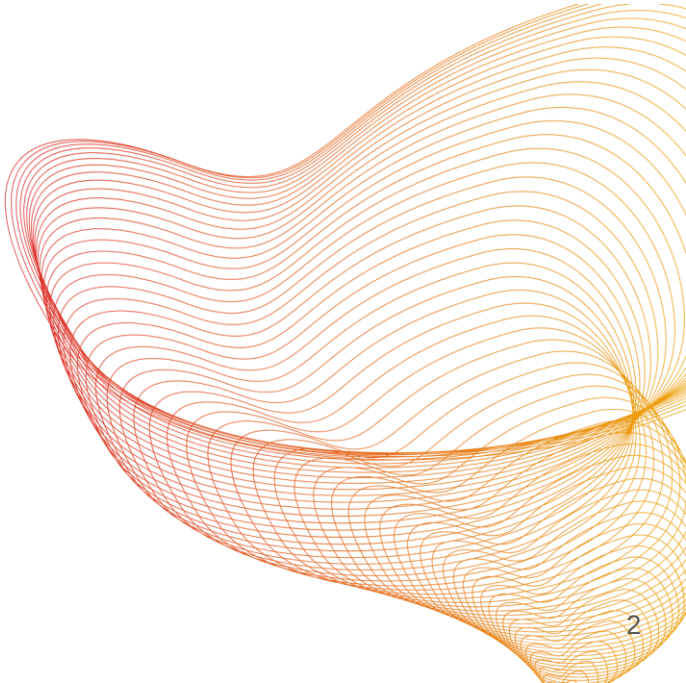
Graphes de connaissances :

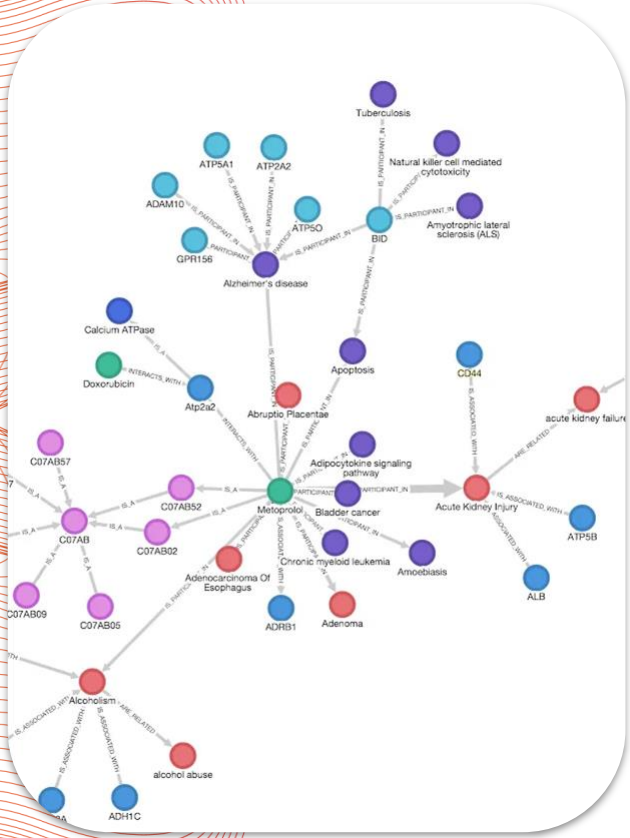
Construction, manipulation et analyse avec l'outil neo4j

Amira MOUAKHER
amira.mouakher@univ-perp.fr

Le plan

- Graphe de connaissances
- Graphe de connaissances vs. ontologie
- Focus sur Neo4j
- Quelques exemples pratiques



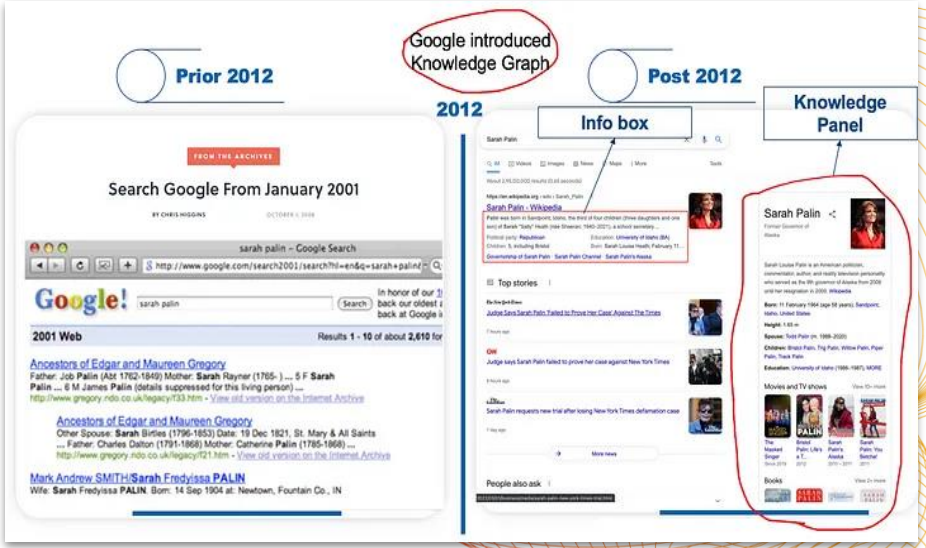


Graphe de connaissances

Qui utilise les graphes de connaissances?

- Fait partie des trois domaines connaissant la croissance la plus rapide à l'échelle mondiale.
- Savez-vous ce que les entreprises suivantes ont en commun ?

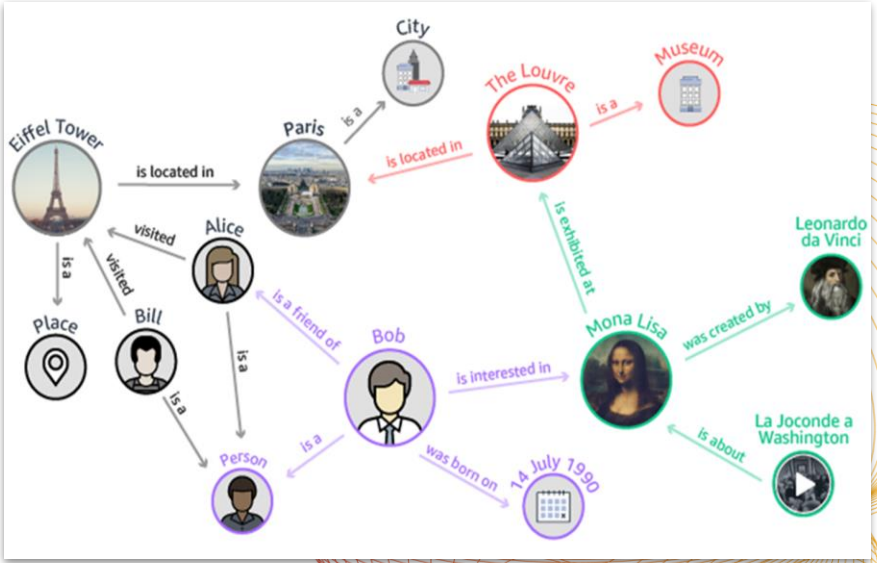
- En 2012, Google a introduit le Google Knowledge Graph pour son moteur de recherche.



- Obtenir une **compréhension plus approfondie** et des **analyses plus puissantes**.

Qu’est ce qu’un graphe de connaissances ?

- **Structure de données** sémantique.
- Représente des informations : nœuds (**entités**) et de liens (**relations**) entre ces entités.
- Les nœuds et liens sont enrichis de **métadonnées** et d'**informations sémantiques.**
- Décrire des faits, des concepts, des entités du monde réel et leurs relations de manière formelle et interconnectée.
- Utilise des standards de représentation sémantique, tels que RDF (Resource Description Framework) et OWL (Web Ontology Language).
- Représente les données dans leur **contexte**, d'une manière **compréhensible** à la fois pour les humains et les machines.

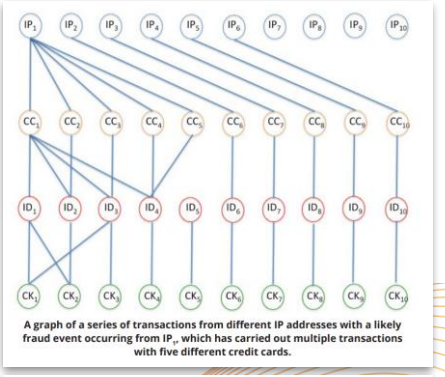


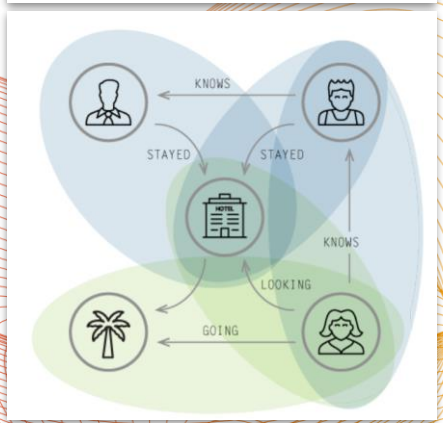
Quelques avantages :

- **Intégration des données :** peuvent intégrer des données provenant de sources hétérogènes, y compris des données structurées et non structurées, et offrir une vue unifiée des données.
- **Facilité de compréhension :** utilisent une représentation visuelle et intuitive des données qui est facile à comprendre et à manipuler. Cela facilite la compréhension des relations même pour les non-experts.
- **Apprentissage automatique et intelligence artificielle :** peuvent être utilisés pour entrainer des modèles d'apprentissage automatique et d'intelligence artificielle en fournissant une vue structurée et interconnectée des données.
- **Personnalisation :** peuvent être utilisés pour personnaliser les expériences utilisateur en proposant des recommandations adaptées au comportement et aux préférences de l'utilisateur.
- **Scalabilité :** peuvent gérer de grandes quantités de données incrémentales.

Domaines d'application

- **Commerce électronique :** personnaliser les recommandations de produits en fonction du comportement et des préférences des clients, améliorant ainsi l'expérience d'achat en ligne.
- **Santé** : découverte de médicaments, la gestion des dossiers médicaux électroniques et la prise de décisions cliniques.
- **Services financiers :** détection de la fraude, l'évaluation des risques et la prédiction des tendances du marché.
- **Gestion des données d'entreprise :** organiser et exploiter les données d'entreprise de manière plus efficace, notamment dans la gestion des connaissances.
- **Éducation :** personnaliser les expériences d'apprentissage des étudiants, d'identifier les lacunes en matière de connaissances et de fournir des ressources d'apprentissage ciblées.



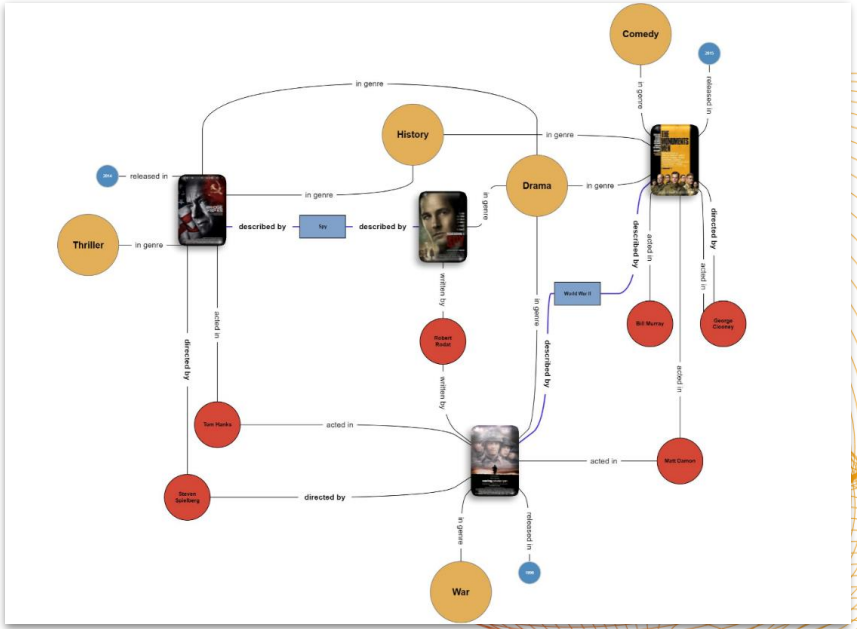


Exposed highly connected networks of offshore tax structures used by the world's richest elites
Analyzing the Panama Papers with Neo4j: Data Models, Queries & More

Exemples de graphes de connaissances en open source

Dataset	Number of Entities	Number of Facts	Number of Classes	Number of Relations	License	Description
DBpedia	4.29 million	411 million	736	2819	CC-BY-SA 3.0	Extracted from Wikipedia, contains entities and relations in various domains
YAGO	5.13 million	1.00 billion	569,751	106	CC-BY-SA 3.0	Extracted from Wikipedia, contains entities and relations in various domains
WordNet	175,979	207,016	4	N/A	Open-source	Lexical database of English words and concepts
Freebase	49.95 million	3.12 billion	53,092	70,902	Open-source	User-contributed facts about people, places, and things
Wikidata	18.69 million	748.53 million	302,280	1874	CC0	Structured data from Wikipedia and other sources
OpenCyc	41,029	2.41 million	116,822	18,028	Commercial and non-commercial	A large ontology and knowledge base of common sense knowledge
IMDb	484 million	N/A	7	N/A	Free for non-commercial use	Contains information about movies, television shows,
MusicBrainz	92.82 million	37.87 million	15	N/A	CC-BY-SA 4.0	Contains information about music artists, albums, songs and more

[Open Source Knowledge Graphs – Diffblog (diffbot.com)]



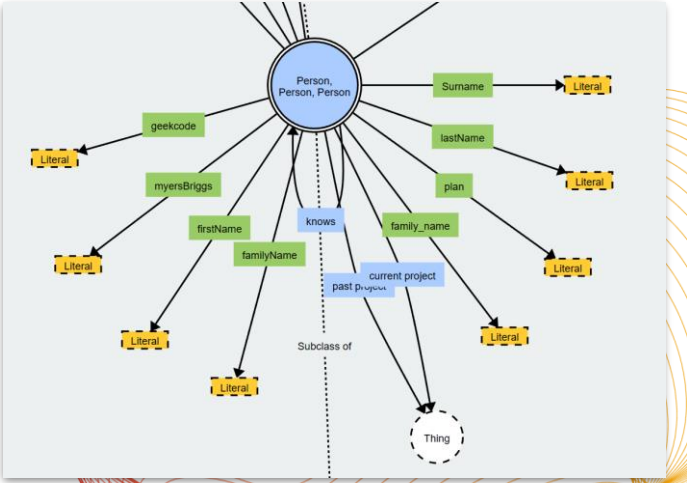
IMDB KG

Graphe de connaissances vs. ontologie

Ontologie et graphe de connaissances

- Deux concepts **liés** mais **distincts** dans le domaine de la gestion des connaissances et de la représentation des données.

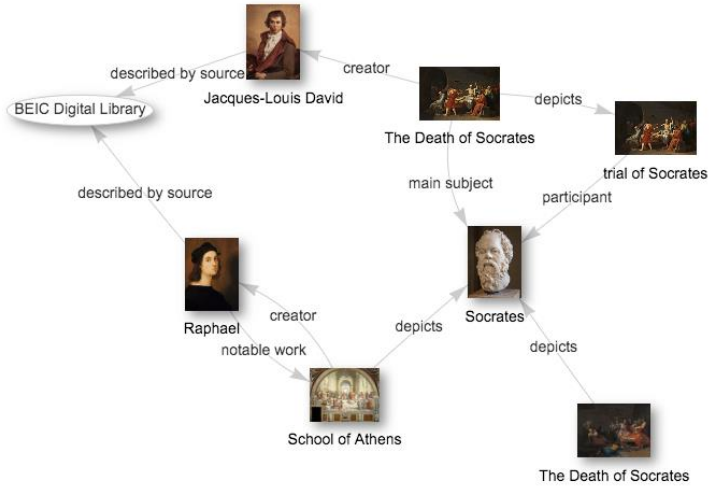
- **Ontologie :**
  - Ensemble formel de concepts, de termes, de relations et de règles qui définissent la structure et la signification des données dans un **domaine particulier**.
  - Etablit une hiérarchie de classes et de sous-classes pour organiser les concepts, leurs propriétés et les relations au sein d'un vocabulaire.
  - Utilisée pour créer une structure sémantique commune et partagée qui permet de comprendre et de raisonner sur les données de manière cohérente.
  - Plus axée sur le schéma (T-BOX) que sur les instances (A-BOX).
  - Représentée en utilisant des langages structurés (OWL ou RDF).



FOAF ontology: describe persons and relations between them FOAF Vocabulary Specification (xmlns.com)

Ontologie et graphe de connaissances

- **Graphe de connaissances :**
  - Structure de données utilisant des nœuds et des arêtes pour représenter des informations et leurs relations de manière interconnectée.
  - Peut être basé sur une ontologie (utilise une structure sémantique définie pour organiser les données).
  - Peut inclure des données non structurées ou semi-structurées, ce qui le rend plus flexible que strictement une ontologie.
  - Utilisé pour modéliser des domaines complexes et permet de naviguer efficacement à travers des données interconnectées.



"Une ontologie peut être considérée comme le schéma de données du graphe de connaissances."

Representation des données (RDF)

⇒ Représentation RDF (Resource Description Framework)

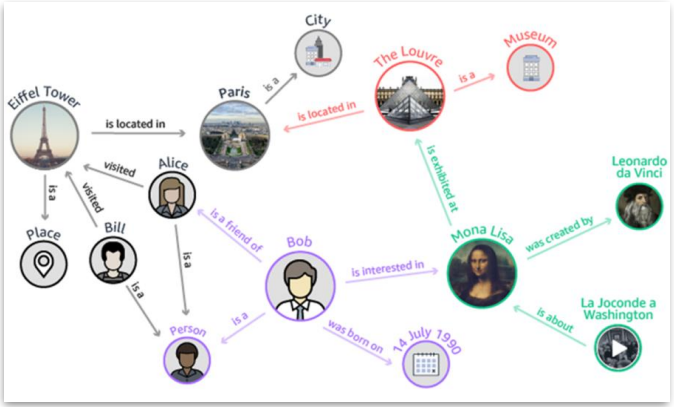
- Norme du World Wide Web Consortium (W3C) pour l'échange de données sur le web.

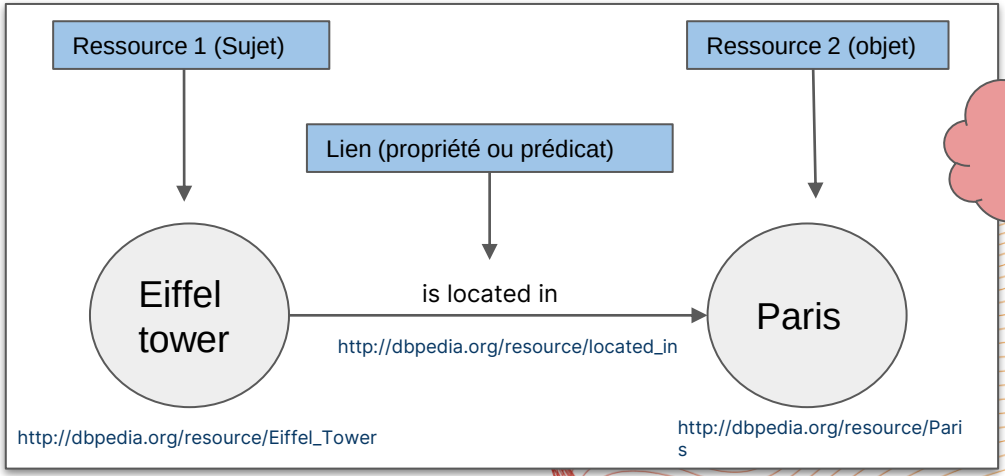
- Représenter des informations sous forme de graphes, où les données sont organisées en **triplets**.
  - **Sujet :** la ressource à décrire.
  - **Prédicat :** un type de propriété applicable à cette ressource.
  - **Objet :** soit des données, soit une autre ressource, et cela constitue la valeur de la propriété.
- Repose sur l'utilisation d'identifiants uniques de ressources (**URI**) pour garantir que les données sont référencées de manière unique sur le web.
- **Les Ontologies** sont basées sur cette représentation.

All Standards and Drafts - W3C

World Wide Web Consortium (W3C)
Current Members & Testimonials (w3.org)

Representation des données (RDF)





C'est un triplet

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dbpedia="http://dbpedia.org/resource/">

  <rdf:Description rdf:about="http://dbpedia.org/resource/Eiffel_Tower">
    <dbpedia:isLocatedIn rdf:resource="http://dbpedia.org/resource/Paris"/>
  </rdf:Description>

</rdf:RDF>
```

Fichier au format **XML/RDF**

Representation des données (RDF)

Stockage avec la représenation RDF

- Bases de données RDF : **Triple Store**
- Opensource ou commercial (prix elevé) DB-Engines Ranking - popularity ranking of RDF stores
- Le "triplestore RDF" (base de données de graphe sémantique) est un type de base de données de graphe qui stocke des faits sémantiques.
- Conçue pour le stockage et la récupération de triplets via des requêtes sémantiques.

The Enterprise Knowledge Graph Platform | Stardog

Ontotext

Amazon Neptune est un service de base de données orientée graphe entièrement géré | Amazon Neptune | Amazon Web Services

Apache Jena - Home

OpenLink Software: Virtuoso Homepage (openlinksw.com)

Representation des données (RDF)

Le langage de requête est **SPARQL**

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/property/>

SELECT ?lat ?long
WHERE {
  <http://dbpedia.org/resource/Paris> dbo:lat ?lat.
  <http://dbpedia.org/resource/Paris> dbo:long ?long.
}
```

Cette requête récupère la latitude et la longitude de Paris sur DBpedia.

Representation des données (LPG)

⇒ Représenation LPG - Label Property Graph Representation

- Les données sont stockées sous forme **d'Entités** et de **Relations**, et les deux ont leurs **propriétés** et leurs **valeurs**.
- Pas de triplet, pas d'URI, pas de norme, propriétaire et commercial.
- Les données sont structurées sous forme de graphe.
- chaque nœud peut être étiqueté avec un ou plusieurs labels.
- Ces étiquettes permettent de catégoriser les nœuds en fonction de leur type ou catégorie.

Representation des données (LPG)

Stockage avec la représentation LPG

- Bases de données graphe.
- Opensource ou commercial (prix elevé) DB-Engines Ranking - popularity ranking of graph DBMS.
- L'utilisation des concepts de la théorie des graphes.
- Permet une modélisation flexible des données.
- Plus intuitif et simple à utiliser.
- Bonne intégration avec les langages de programmation.
- Solution pour l'industrie.

neo4j https://neo4j.com

TigerGraph Graph Analytics Platform | Graph Database | TigerGraph

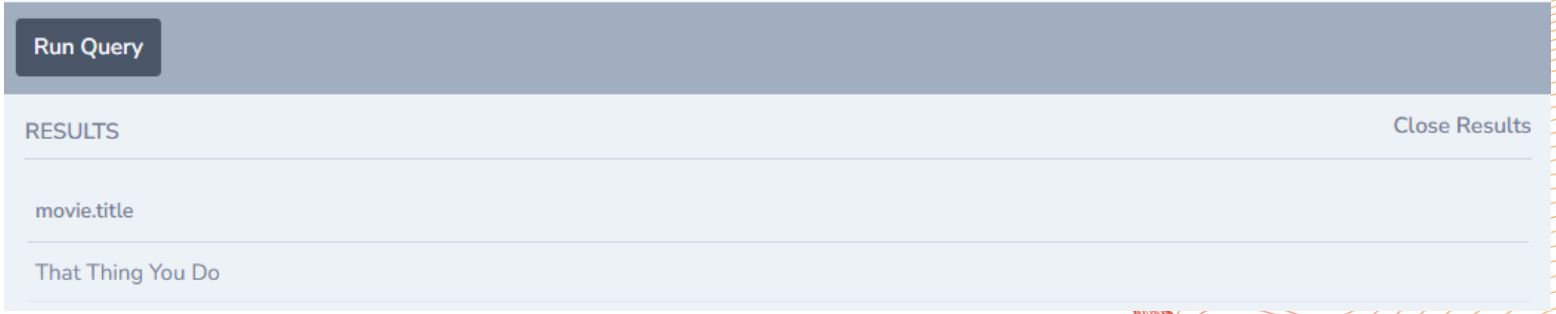
ORACLE Graph Database | Oracle

ArangoDB Message from ArangoDB

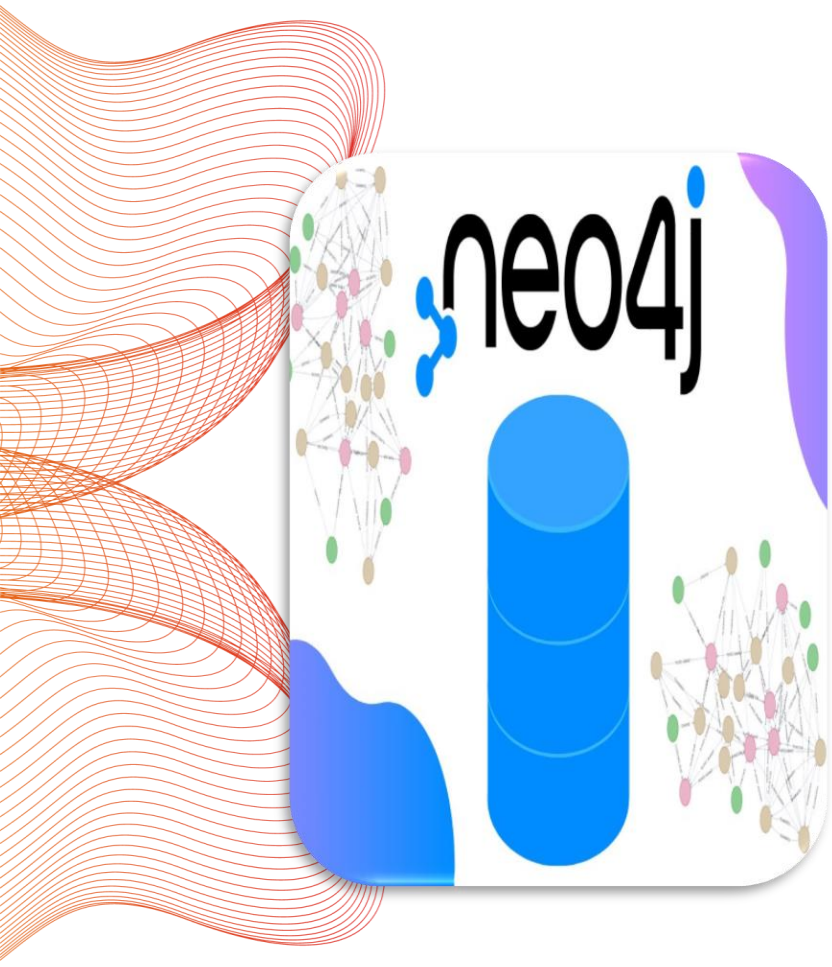
Representation des données (LPG)

Le langage de requête **(utilisation de CYPHER)**

```
MATCH (:Person {name: 'Tom Hanks'})-[:DIRECTED]->(movie:Movie)
RETURN movie.title
```



Trouvez les films que Tom Hanks a réalisés et retourner uniquement le titre du film.

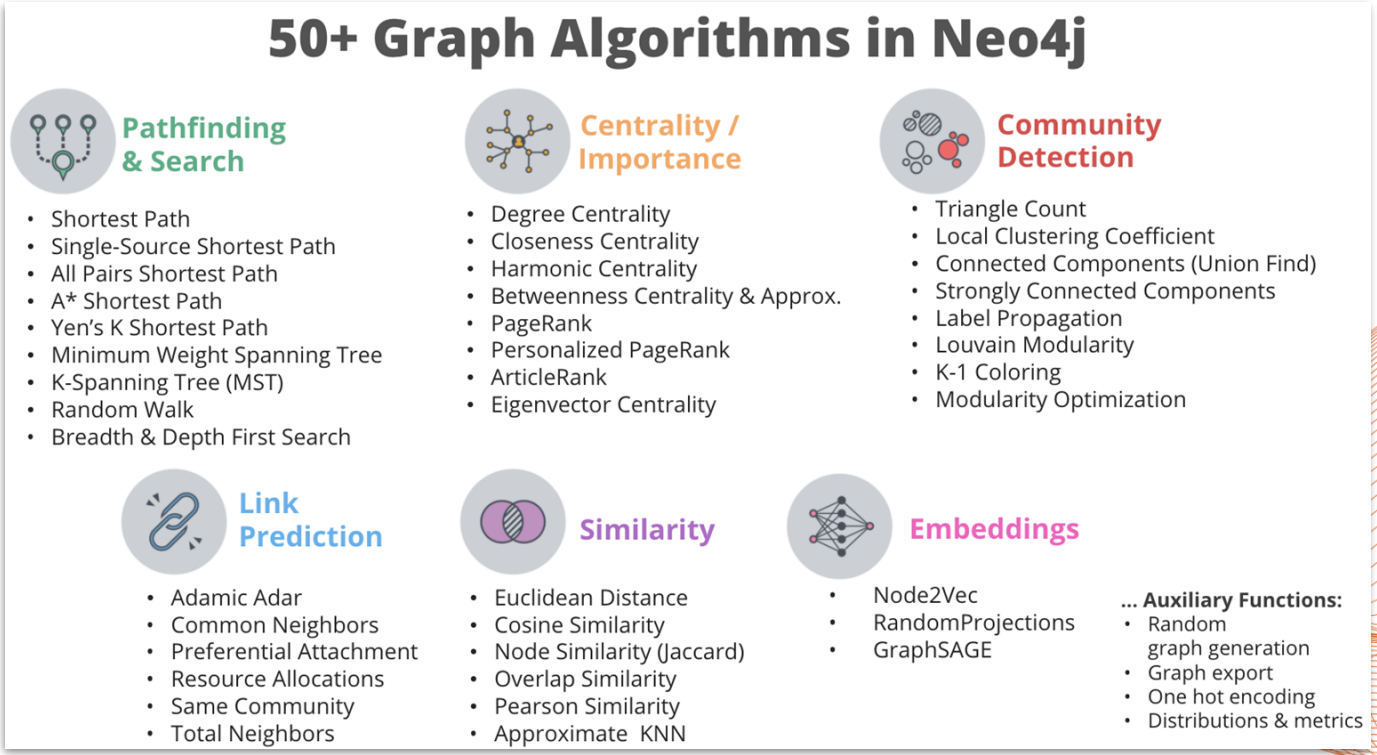


Focus sur Neo4j

Focus sur Neo4j

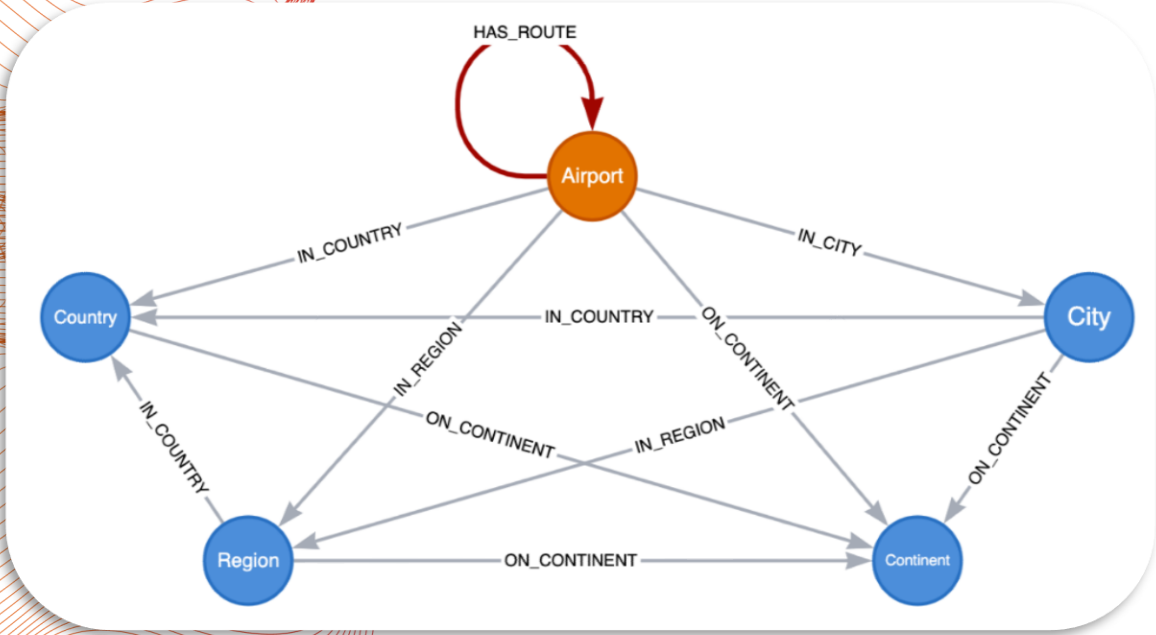
- Base de données graphe.
- Entièrement transactionnel (ACID).
- Grande flexibilité en termes de modélisation des données.
- Idéal pour les données fortement connectées avec des relations complexes.
- Extrêmement rapide lorsqu'il s'agit de consulter des données connectées.
- Hautement évolutif, jusqu'à plusieurs milliards de nœuds/relations/propriétés.
- Version cloud, docker, desktop.
- Plusieurs plugins (NeoSemantics, Bloom, APOC, etc).

Graph Data science dans Neo4j



[Graph algorithms - Neo4j Graph Data Science]

Quelques exemples pratiques



- Dans la suite, nous allons passer à la partie pratique :

1. Connecter Neo4j avec Jupiter Notebook.
2. Manipuler la base « Airport routes » à travers des requêtes Cypher.
3. Appliquer l'algorithme Page Rank sur le graphe.
4. Importer un fichier RDF avec le plugin Neosemantics.

Merci pour votre attention 😊