Université Sidi Mohamed Ben Abdellah
Faculté des Sciences Dhar El Mahraz
Fès

Cours Big Data

Master Recherche en Informatique Décisionnelle et Vision Intelligente (MIDVI)

Préparé par :
Pr. Noura AHERRAHROU

Du SGBD SQL au SGBD NoSQL

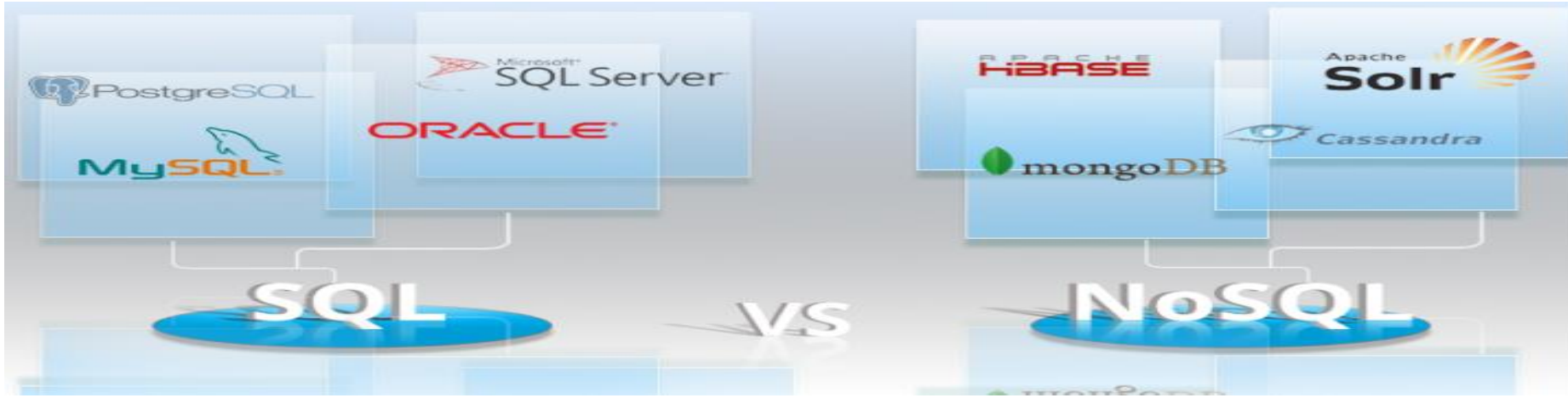
En effet, avec l'explosion de l'information, les entreprises ont de plus en plus de mal à gérer des données qui arrivent sous des formes de plus en plus variées et qui sont produites de plus en plus rapidement. Les géants du Web ont très tôt ressenti ce problème et le besoin pressant de gérer efficacement ce flux important de données.

Du SGBD SQL au SGBD NoSQL

- L'approche traditionnelle consiste à **centraliser le stockage** et l'exploitation des données sur un serveur de SGBDR. Cependant, avec l'explosion phénoménale des données, les SGBDR ont montré très rapidement leurs limites face, d'une part, à la forte volumétrie des données, et d'autre part à la diversité des types de données.
- En effet, les SGBDR sont conçus pour gérer uniquement des **données structurées** (tabulaires). De plus, l'augmentation du volume des données **accroît le temps de latence des requêtes**. Cette latence est préjudiciable dans le cadre de nombreux métiers nécessitant des réponses en temps quasi réel.

Du SGBD SQL au SGBD NoSQL

- Dès lors, **la solution** n'est plus de centraliser la gestion des données sur ce seul serveur (SGBDR), mais de **distribuer leur stockage et leur requêtage sur plusieurs machines** (un cluster d'ordinateurs).
- Or, les SGBDR ne sont pas par essence des systèmes distribués. C'est pour répondre à ces nouvelles exigences de montée en charge, de disponibilité et de distribution du stockage que les SGBD dits « NoSQL » ont émergé.

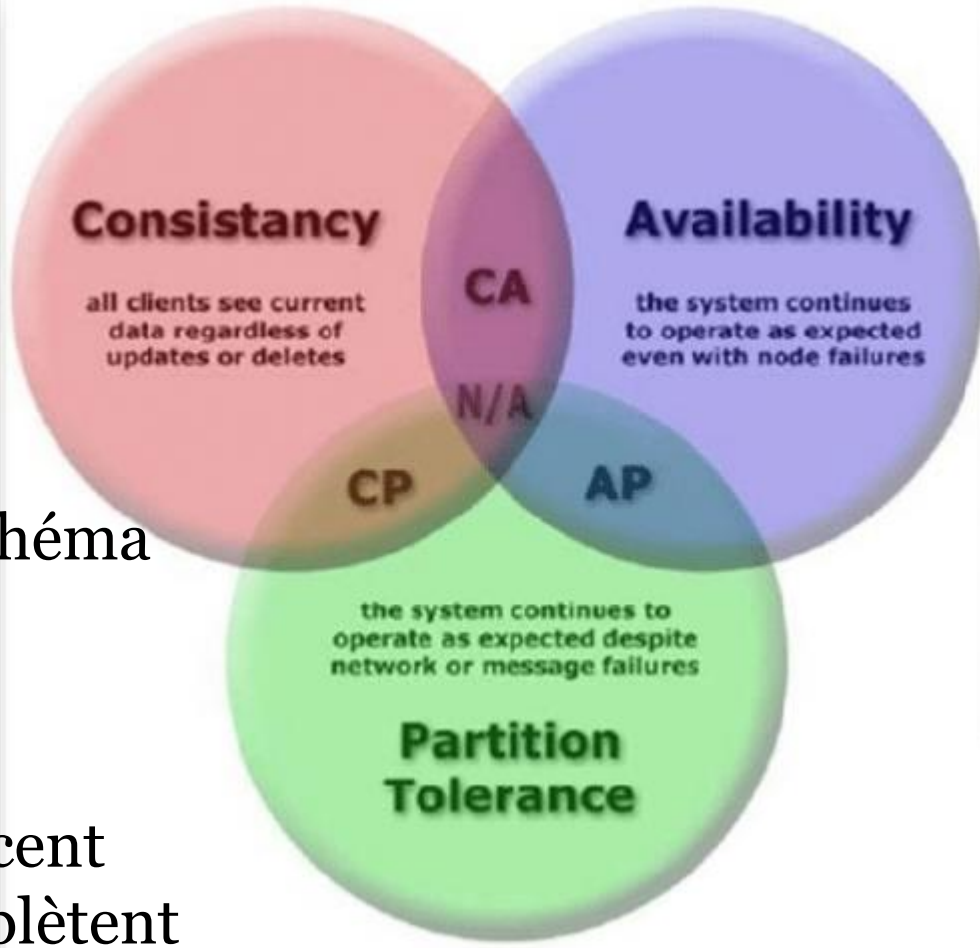


Les bases de données NoSQL

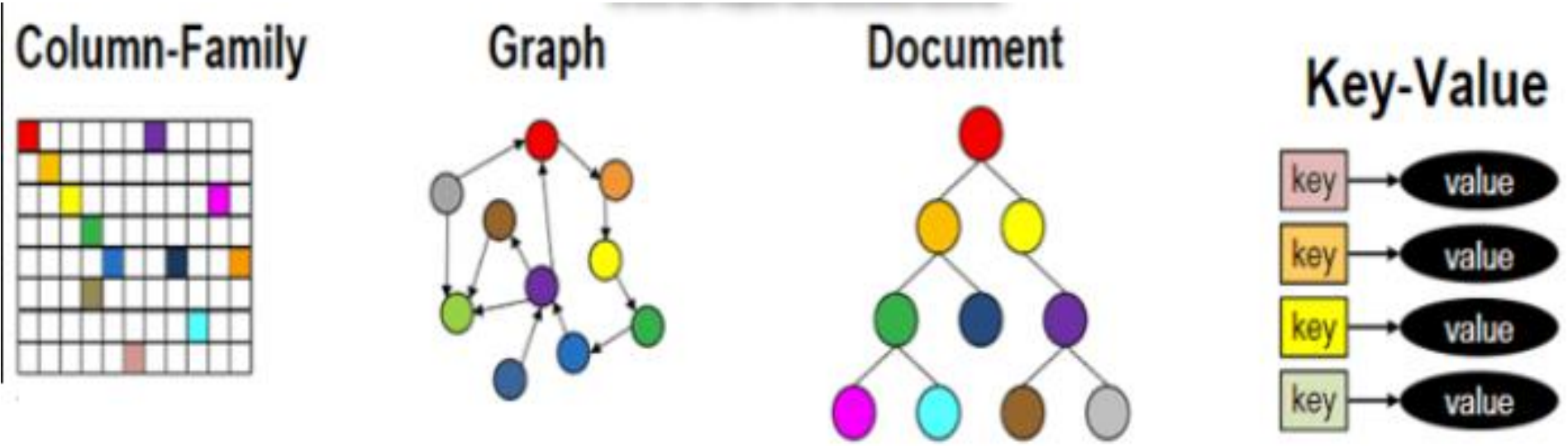
- Beaucoup traduisent NoSQL par « No SQL », d'autres par « Not Only SQL », pour faire référence à des SGBD qui n'utilisent pas (ou presque) le SQL. En réalité, le débat n'est pas là. Le terme « NoSQL » n'a rien avoir avec la présence ou l'absence du SQL dans le SGBD. Il renvoie plutôt à un changement d'approche dans la conception du système et de la base de données (passage de l'approche relationnelle à l'approche non relationnelle).

Les Bases de données NoSQL

- **Principaux atouts**
  - Évolutivité
  - Disponibilité
  - Tolérance aux pannes
- **Caractéristiques**
  - Architecture distribuée
  - Modèle de données sans schéma
  - Utilisation de langages et interfaces qui ne sont pas uniquement du SQL
  - Les SGBD NoSQL ne replacent pas les SGBDR mais les complètent en palliant leurs faiblesses



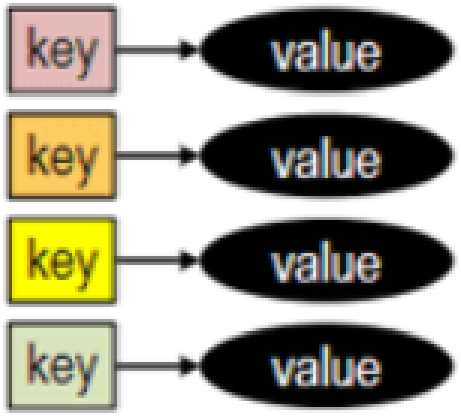
Panorama des SGBD NoSQL



La relation (ou table) n'est pas adaptée au contexte du big data. En réponse à cette limite, de nouvelles approches de stockage, plus souples, ont été élaborées et ont donné naissance à quatre catégories de SGBD NoSQL : **orientés clé/valeur**, **orientés colonnes**, **orientés documents**, **orientés graphes**.

Les SGBD orientés clé/valeur

- Le type le plus élémentaire de base de données NoSQL.
- Conçues pour sauvegarder les données sans définir de schéma
- Toutes les données sont sous forme de clef/valeur
  - La valeur peut être une chaîne de caractères, un objet,...
  - La donnée est opaque au système: il n'est pas possible d'y accéder sans passer par la clef



- Communications se résumant surtout aux opérations PUT (ajout d'un pointeur), GET (obtention d'un objet à partir de son pointeur) et DELETE (suppression d'un pointeur et de l'objet associé).

Les SGBD Clé/valeur

Cas d'utilisation

- Les moteurs de recherche tels que Google utilisent ces SGBD pour stocker simplement des sites web entiers sous la forme clé/valeur.
- Amazon utilise aussi ce type de système pour gérer le stockage des images et vidéos dans son offre Cloud

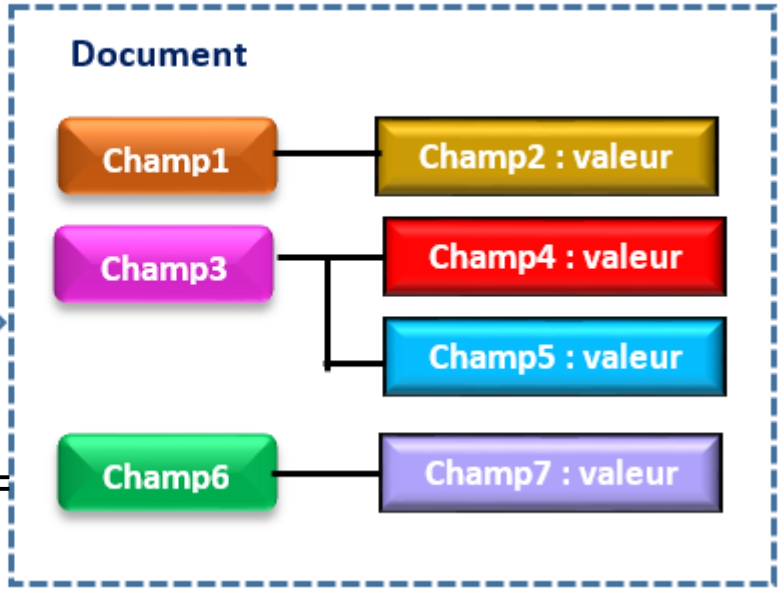
Logiciels

- Amazon Dynamo (Riak est l'implémentation open source).
- Redis (projet sponsorisé par VMWare).
- Oracle NoSQL Database

Amazon DynamoDB (Beta) ORACLE BERKELEY DB 11g redis

Les SGBD orientés documents

- Ce sont des SGBD clé/valeur, à la différence que les valeurs sur lesquelles pointent les clés sont des documents (JSON le plus souvent, ou XML).

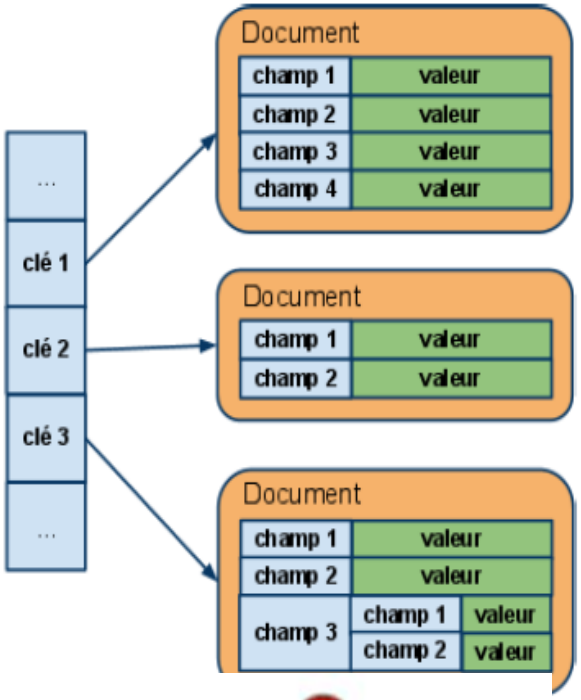


- Document (structure arborescente)= collection de couples (clé, valeur)
- Valeur de type simple, ou composée de plusieurs couples (clé, valeur)
- Les documents ne sont pas généralement forcés d'avoir un schéma. Ils sont donc flexibles et faciles à modifier.
- Pouvoir de récupérer, via une seule clé, un ensemble d'informations structurées de manière hiérarchique.

Les SGBD orientés documents

Cas d'utilisation

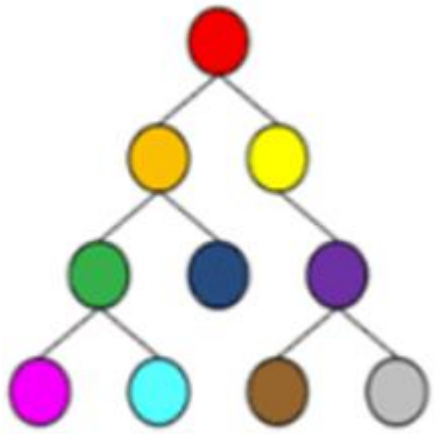
- outils de gestion de contenu (Content Management System(CMS)), catalogues de produits,web analytique, analyse temps réel, enregistrement d'événements, stockage de profils utilisateurs, systèmes d'exploitation, gestion de données semi-structurées



Logiciels

- CouchDB, RavenDB, MongoDB, Terrastore

mongoDB RAVEN DB terrastore terrastore CouchDB relax



Les SGBD orientés documents

Exemple

- Un document JSON pourrait, par exemple, prendre toutes les données stockées dans une ligne qui s'étend sur 20 tables d'une base de données relationnelle et de les regrouper dans un seul document/objet.

Les SGBD orientés colonnes

Stockage colonne

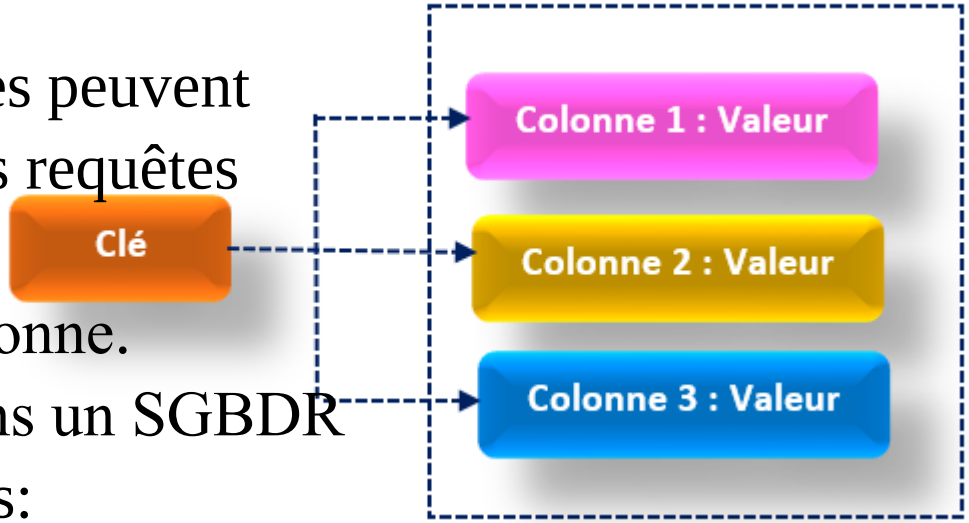
a1	a2	a3	b1	b2	b3	c1	c2	c3

a	b	c
a1	b1	c1
a2	b2	c2
a3	b3	c3
a4	b4	c4
a5	b5	c5

- Données stockées en colonnes.
- C'est une évolution de la BD clé/valeur.
- La colonne est l'entité de base représentant un champ de donnée, chaque colonne est définie par un couple (clé, valeur) avec une est ampille(pour gérer les versions et les conflits)
- Une super-colonne est une colonne contenant d'autres colonnes
- Une famille de colonnes regroupe plusieurs colonnes ou supercolonnes où les colonnes sont regroupées par ligne et chaque ligne est identifiée par un identifiant unique et par un nom unique

Les SGBD orientés colonnes

- Les stockages orientés colonnes peuvent améliorer les performances des requêtes car ils peuvent accéder à des données spécifiques d'une colonne.
- Modèle proche d'une table dans un SGBDR mais ici le nombre de colonnes:
  - Est dynamique.
  - Peut varier d'un enregistrement à un autre ce qui évite de retrouver des colonnes ayant des valeurs NULL.



Les SGBD orientés colonnes

Cas d'utilisation

- Analyse de données, traitement analytique en ligne (OnLine Analytical Processing(OLAP)), exploration de données(data mining),entrepôt de données(data warehouse), gestion de données semi-structurées, jeux de données scientifiques, génomique fonctionnelle, journalisation d'événements et de compteurs, analyses de clientèle et recommandation, stockage de listes (messages, posts, commentaires,…), traitements massifs.

Logiciels

- BigTable, HBase, Cassandra, SimpleDB

Les SGBD orientés colonnes

Exemple

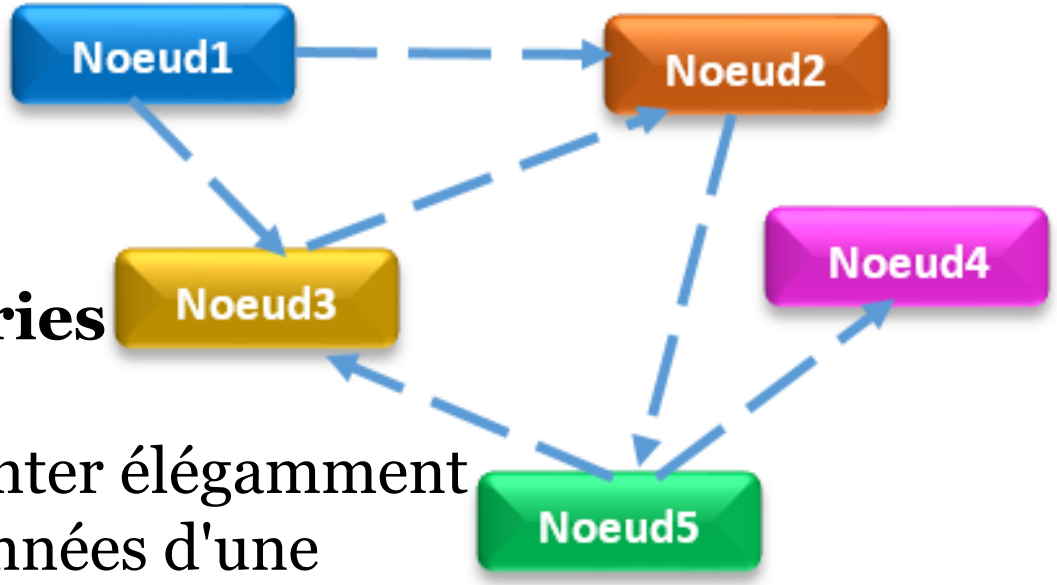
« Orientée ligne »

Id	Nom	Prénom
1	Brico	Juda
2	Diote	Kelly

« Orientée Colonne »

Ligne « row »	Colonne « column »	Valeur « value »
1	Nom	Brico
1	Prénom	Juda
2	Nom	Diote
2	Prénom	Kelly

Les SGBD orientés graphes



- Une base de données de type graphe stocke les données dans un **graphe**.
- Elle est basée sur les **théories des graphes**.
- Elle est capable de représenter élégamment n'importe quel type de données d'une manière **hautement accessible.**
- La gestion d'un graphe (apriori orienté)c.-à-d. la modélisation, le stockage et la manipulation de données complexes liées par des relations non-triviales ou variables
- Chaque **nœud** représente une entité (comme un étudiant ou une entreprise) et chaque **arc** représente un lien ou relation entre deux noeuds.

Les SGBD orientés graphes

- **Quand le nombre de nœuds augmente, le coût d'une étape local(ou hop) reste le même.**
- **Conçues pour les données dont les relations sont représentées comme graphes,et ayant des éléments interconnectés, avec un nombre indéterminé de relations entre elles.**
- Adapté aux traitements des données des réseaux sociaux

Les SGBD orientés graphes

Cas d'utilisation

- Moteurs de recommandation, informatique décisionnelle, web sémantique, internet des objets(internet of things (IoT)), sciences de la vie et calcul scientifique(bioinformatique,...), données géospatiales, données liées, données hiérarchiques(catalogue des produits, généalogie,...), réseaux sociaux, réseaux de transport, services de routage et d'expédition,services financiers (chaîne de financement, dépendances, gestion des risques, détection des fraudes,...),données ouvertes(opendata)

Logiciels

- Neo4J, OrientDB,Titan

Les SGBD orientés graphes

Exemple