BDBIO - Introduction aux SGBD non relationnels

Fabien Duchateau

fabien.duchateau [at] univ-lyon1.fr

Université Claude Bernard Lyon 1

2024 - 2025

https://perso.liris.cnrs.fr/fabien.duchateau/BDBIO/

Rappels

Les données d'application sont principalement gérées par des systèmes de gestion de bases de données (SGBD), qui suivent un modèle de données

Modèle	**Sérialisation**	**LDD**	**LMD**	**Exemples SGBD**
Relationnel	n-uplets	SQL	SQL	SQLite, MariaDB, PostgreSQL

Avec l'émergence d'Internet et du Big Data, se pose le problème du passage à l'échelle et des limites du modèle relationnel

Le Big Data

Big Data : modélisation, stockage et analyse d'un ensemble de données volumineuses, croissantes et hétérogènes, dont l'exploitation permet la prise de décisions ou la découverte de nouvelles connaissances

Les "3V", caractéristiques du Big Data :

- **Volume** (e.g., plusieurs zettaoctets/an générés sur le web)
- **Vélocité** ou fréquence de génération des données, (e.g., 4000 To/jour pour Facebook en 2016 ou 7000 To/seconde pour le radiotélescope Square Kilometre Array)
- **Variété** ou hétérogénéité (e.g., images, texte, données géo-démographiques)

Extension à "5V" avec véracité (provenance) et valeur (ajoutée)

https://fr.wikipedia.org/wiki/Big_data

Exemples d'application du Big Data

- Création de cartes de navigation par Fontaine Maury, au $19^{\text{ème}}$ siècle, à partir de vieux journaux de bord (précurseur)
- Large Hadron Collider (LHC), un accélérateur de particules
- Décodage du génome humain
- Programmes de surveillance (e.g., Prism)
- Réduction de 22% du gaspillage alimentaire

- Découverte d'un effet secondaire dû à la prise de deux médicaments par analyse des requêtes des internautes (Yahoo)
- Confirmation de la censure par analyse de la fréquence de noms de personnes (Chagall en Allemagne, Trotski en URSS)

Motivation

Distribution des données sur différents serveurs et data centers :

- Passage d'un système vertical ("dopage" du serveur) à un système horizontal (ajout de machines "basiques")
- Contraignant avec des SGBD relationnels (transaction, jointure, contrainte d'intégrité - *détails dans l'UE GGMD*)

Nouveaux besoins :

- Passage à l'échelle (meilleures performances)
- Forte disponibilité, résistance aux pannes
- Gestion flexible des données (schémas dynamiques)

Motivation

Distribution des données sur différents serveurs et data centers :

- Passage d'un système vertical ("dopage" du serveur) à un système horizontal (ajout de machines "basiques")
- Contraignant avec des SGBD relationnels (transaction, jointure, contrainte d'intégrité - *détails dans l'UE GGMD*)

Nouveaux besoins :

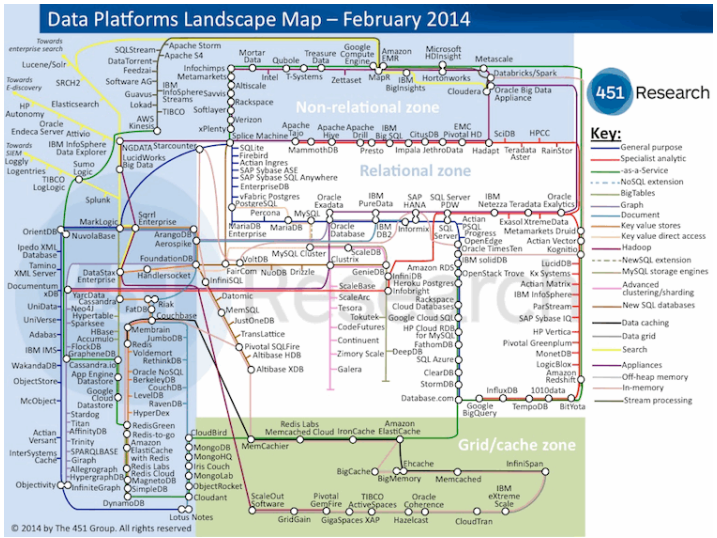
- Passage à l'échelle (meilleures performances)
- Forte disponibilité, résistance aux pannes
- Gestion flexible des données (schémas dynamiques)

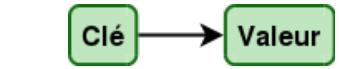
Mouvement NoSQL, NotOnlySQL, Non Relationnel, NewSQL, ...

Nouveaux modèles post-relationnels

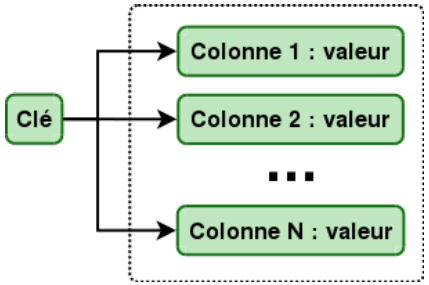
SGBD non-relationnel ≈ entrepôt clé-valeur fortement optimisé

- Détails dans l'UE *MIF24 - BD NoSQL*

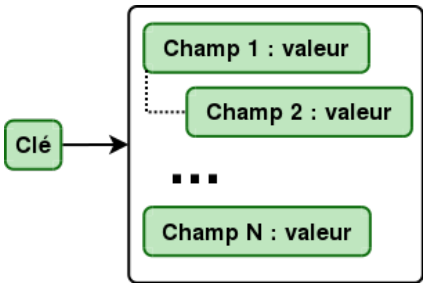




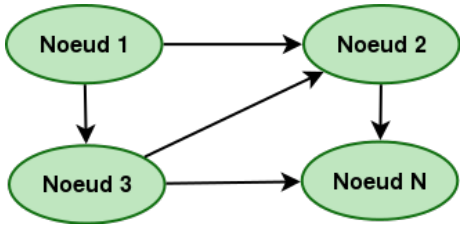
Modèle clé-valeur



Modèle colonne



Modèle document



Modèle graphe

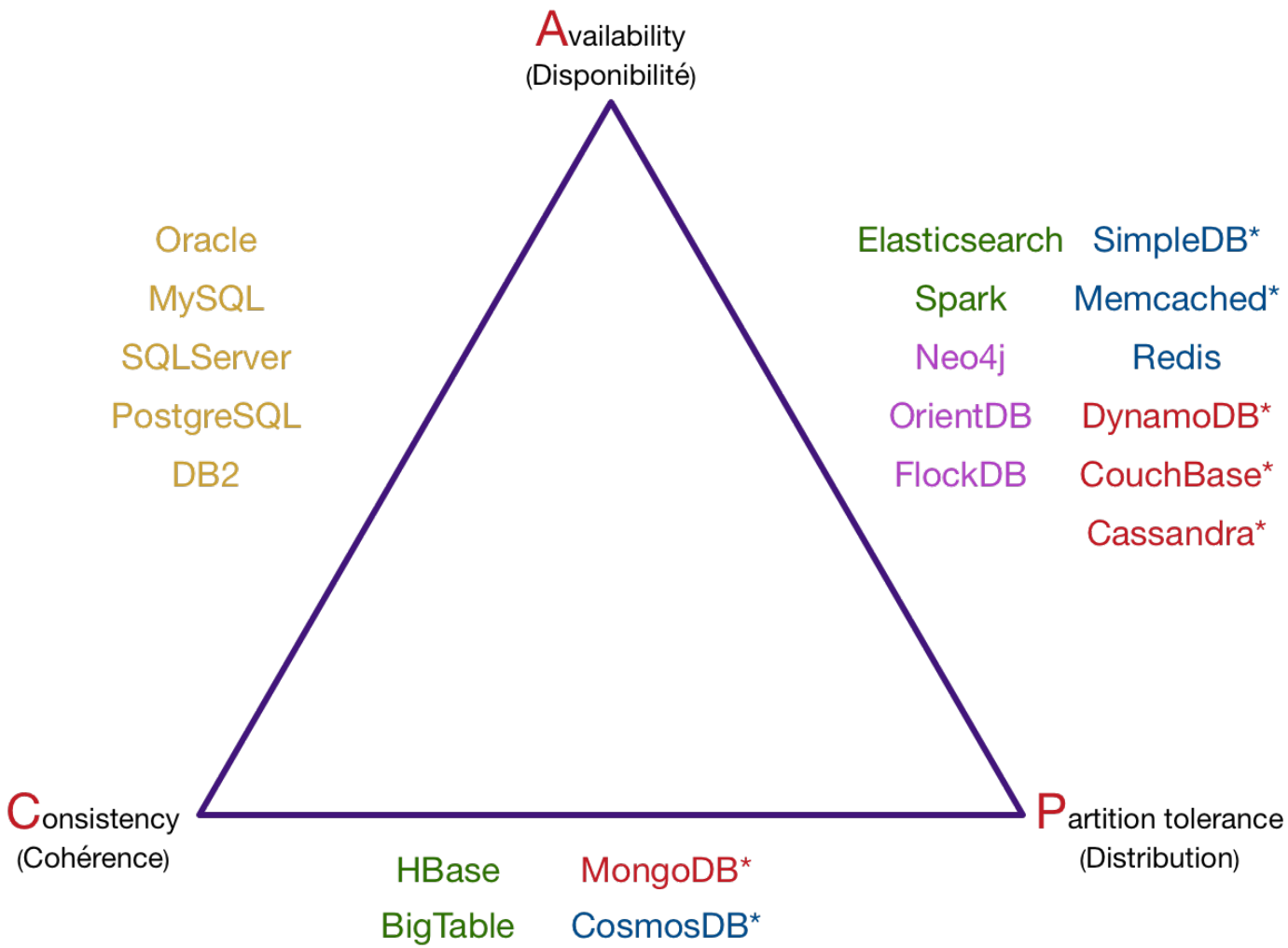
Davoudian et al. A Survey on NoSQL Stores, CSUR (2018)
https://dbdb.io/, https://db-engines.com/

Classification selon le théorème de CAP

Modèle

Relationnel
Orienté Clé-Valeur
Orienté Colonne
Orienté Document
Orienté Graphe

* Possibilité de changer
le mode de cohérence
Cohérence <-> Disponibilité



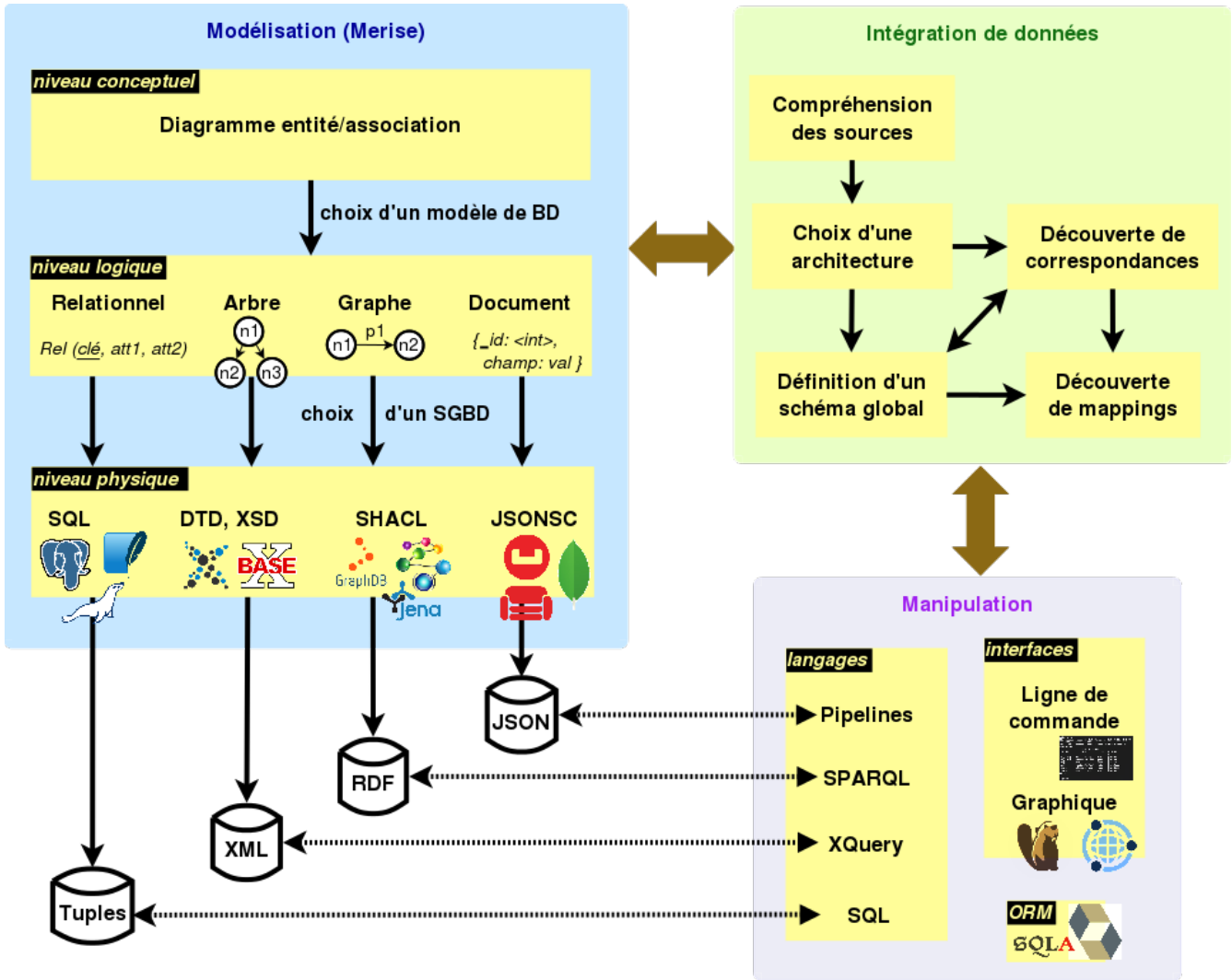
Outils de transformation

De nombreux modèles, mais aussi des outils pour faciliter le passage d'un modèle à un autre, par exemple :

- DuckDB, un SGBD embarqué pour l'analyse de données
  - import de données SQL, CSV, Parquet, JSON, etc.
  - interrogation avec le langage SQL
  - bonnes performances malgré installation locale

- LinkedML, un langage de modélisation pour les données liées
  - définition d'un schéma en YAML
  - transformation d'un schéma en JSON schema, SQL, RDF, SHACL, classes Python, etc.
  - implémentations pour la biologie (BioLink, CRDC, etc.)

---

https://duckdb.org/
https://linkml.io/
https://biolink.github.io/

Aperçu des notions de BDBIO