



INSTITUT  
DE LA  
communication



# Lacs de données et métadonnées

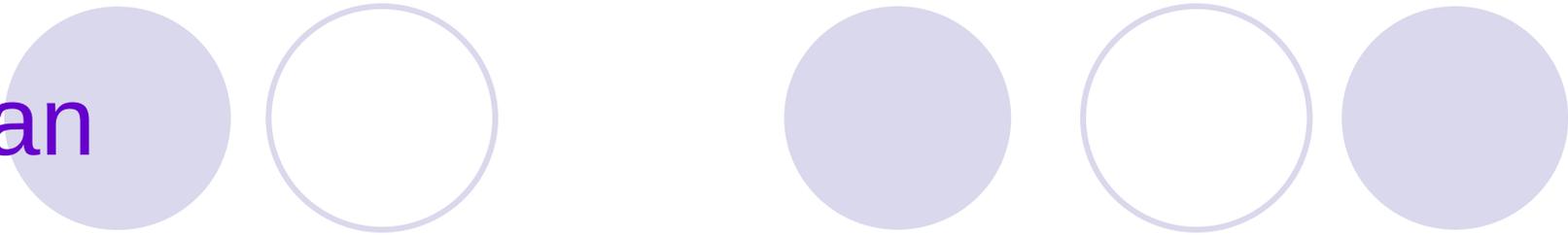
Séminaire Master 2 IDSM-Kharkiv

Année 2024-2025

Jérôme Darmont

<https://eric.univ-lyon2.fr/jdarmont/>

# Plan



- Définitions
- Entrepôts et lacs de données
- Métadonnées et modèles de métadonnées
- Architectures et technologies pour les lacs
- Discussion, travaux de recherche

# Définition de James Dixon (2010)

“If you think of a **datamart as a store of bottled water** – cleansed and packaged and structured for easy consumption – **the data lake is a large body of water in a more natural state.**”

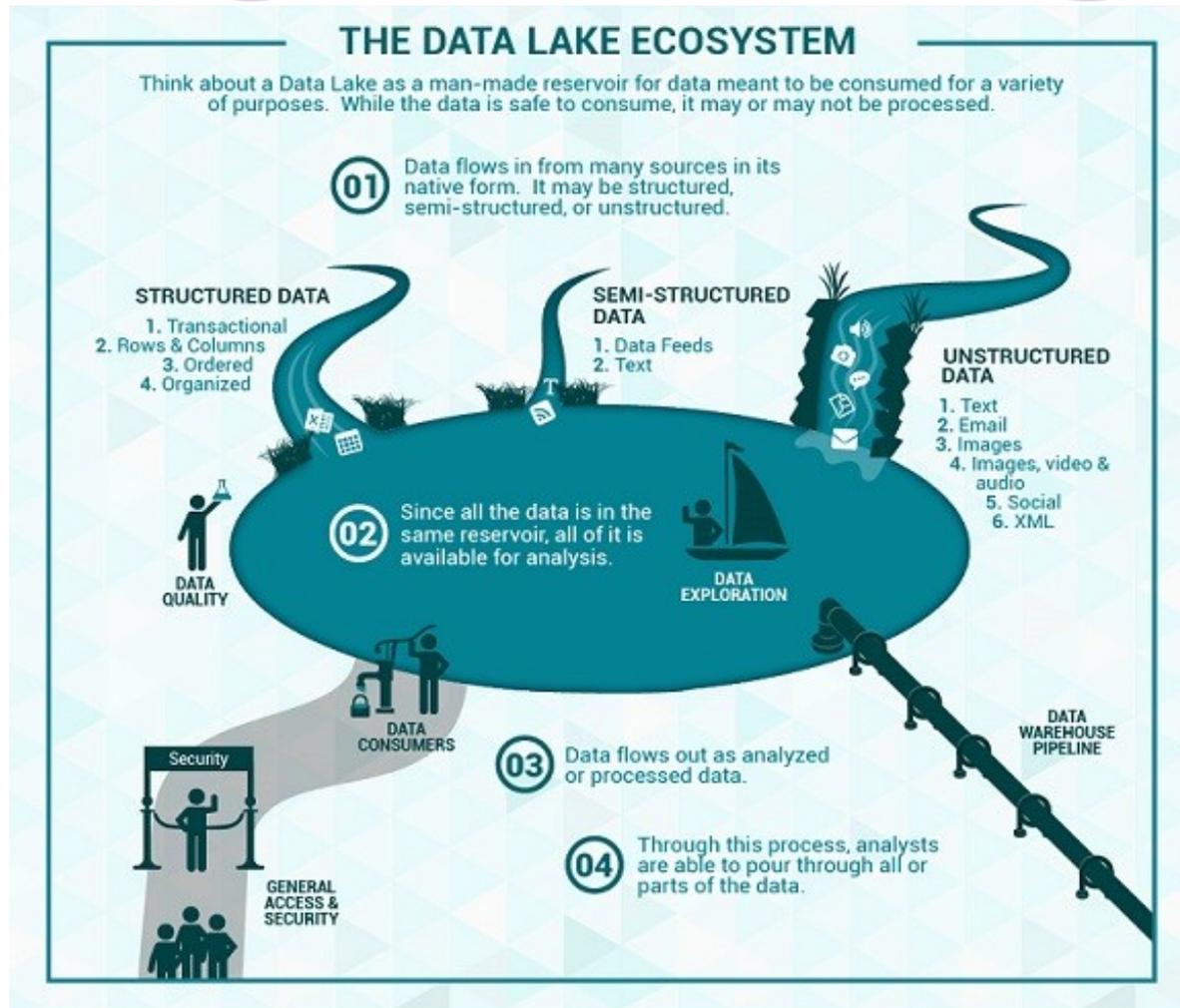
The contents of the data lake **stream in from a source** to fill the lake, and **various users** of the lake come to examine, dive in, or take samples.”



csgsolutions.com

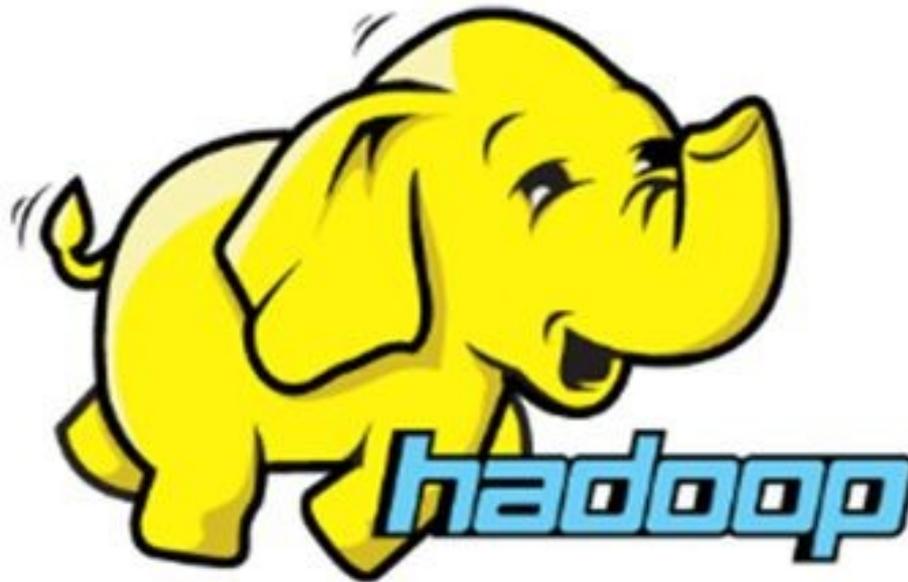
# Plus en détail

dunnsolutions.com

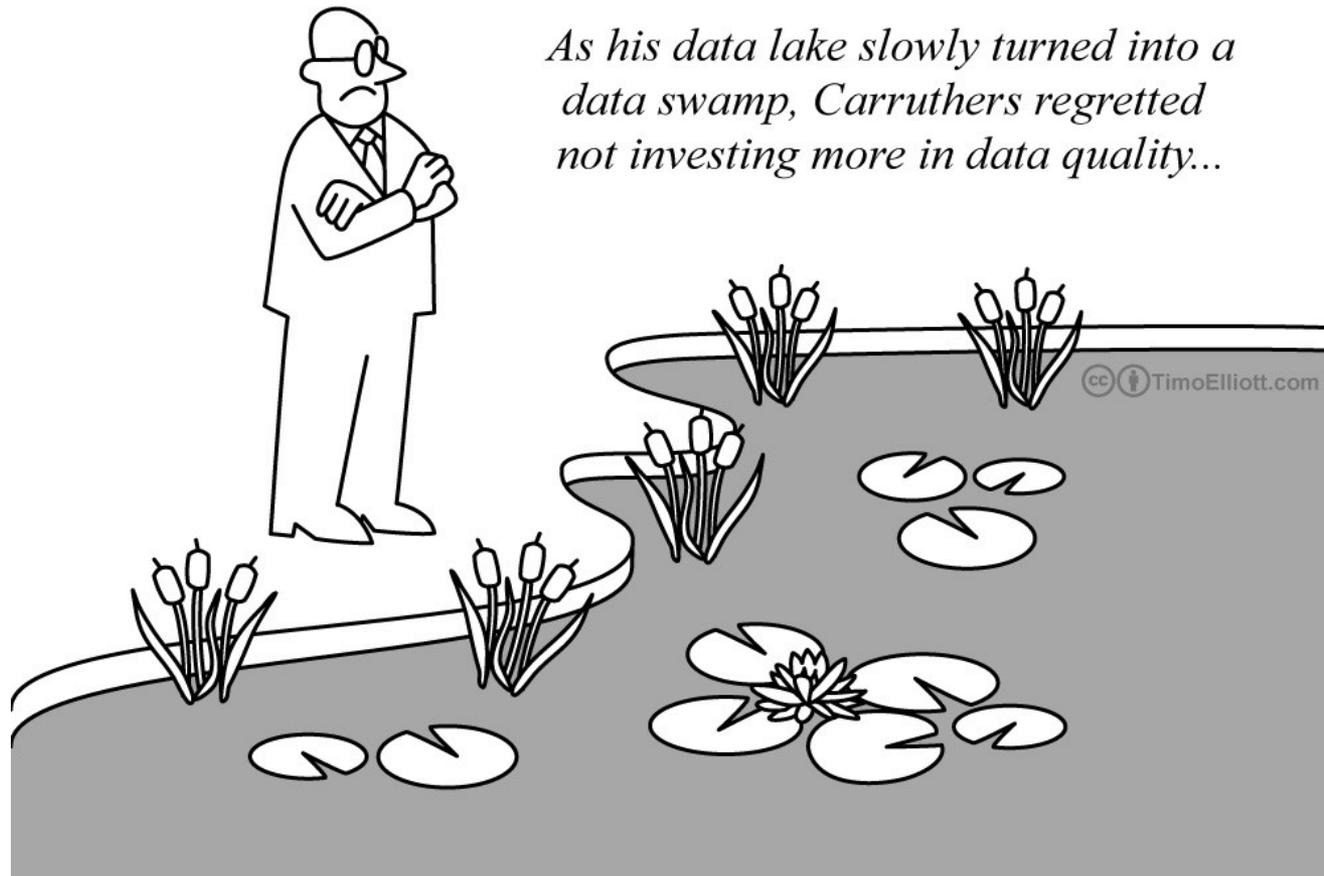


# Définition initiale de l'industrie

...et de quelques auteurs académiques (minoritaire aujourd'hui)



# Danger 1 : le *data swamp*



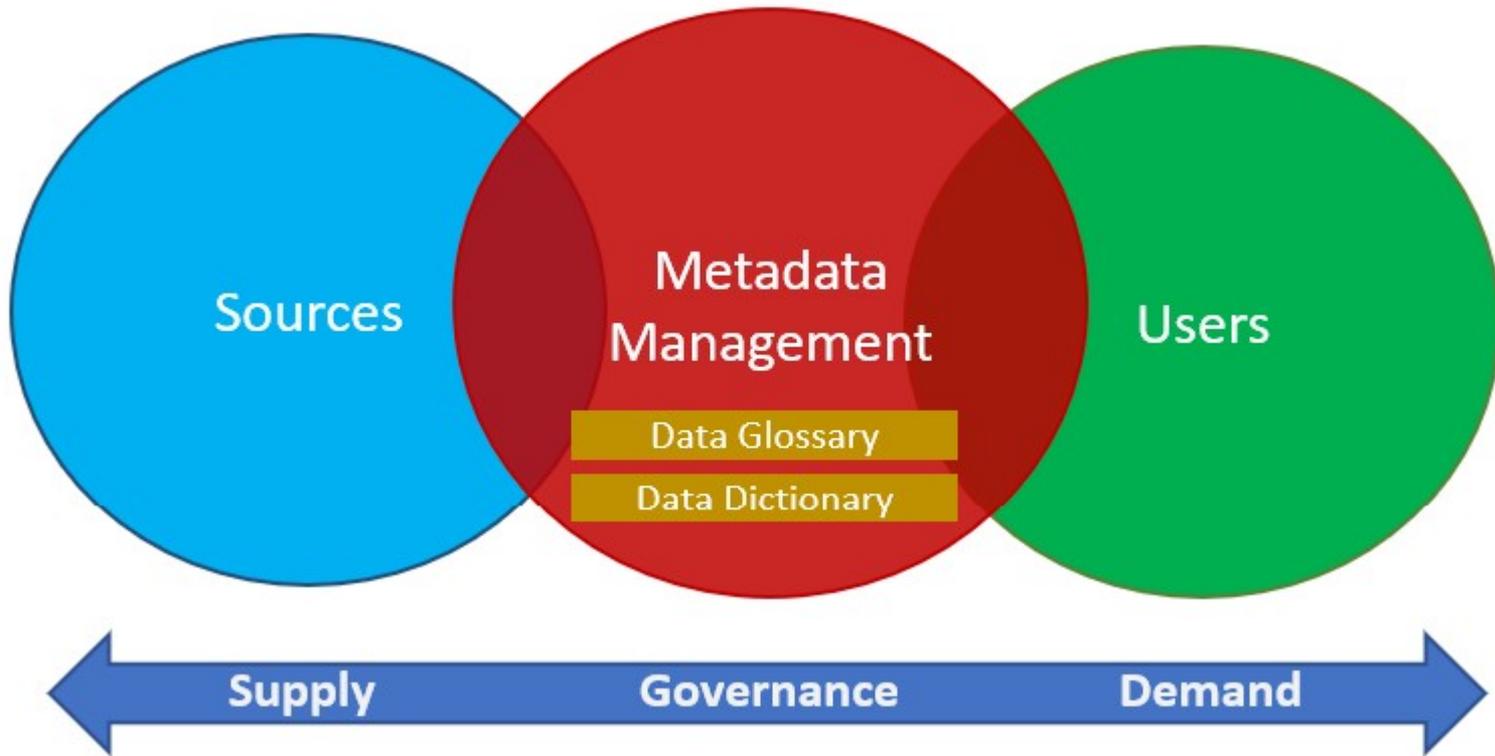
timoelliott.com

# Danger 2 : le *data dump*



[www.cartoonstock.com](http://www.cartoonstock.com)

# Comment éviter ces dangers ?



[medium.com](https://medium.com)

+ cycle de vie des (méta)données

# Définition de Scholly et al. (2019)

## Lac de données

Système **évolutif** (passage à l'échelle) de stockage et d'analyse

Données de tous types dans leur **format natif**

Utilisé **pas seulement** par des *data scientists* et *data analysts*

(utilisateur·trices métiers, chercheur·es...)

# Définition de Scholly et al. (2019)

Caractéristiques des lacs de données

**Catalogue de métadonnées** (qualité des données)

Politique et outils de **gouvernance des données**

Ouverture à **tous types d'utilisateur·trices**

Intégration de **tous types de données**

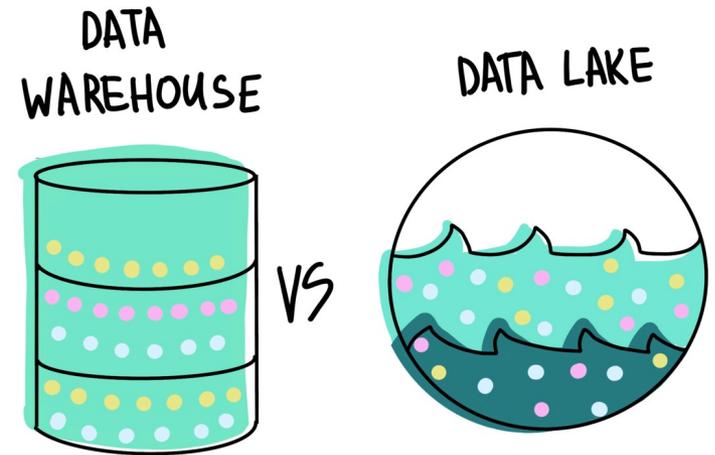
Organisation **conceptuelle, logique et physique**

**Passage à l'échelle** (stockage et traitement)

# Plan

## ✓ Définitions

- Entrepôts et lacs de données
- Métadonnées et modèles de métadonnées
- Architectures et technologies pour les lacs
- Discussion, travaux de recherche

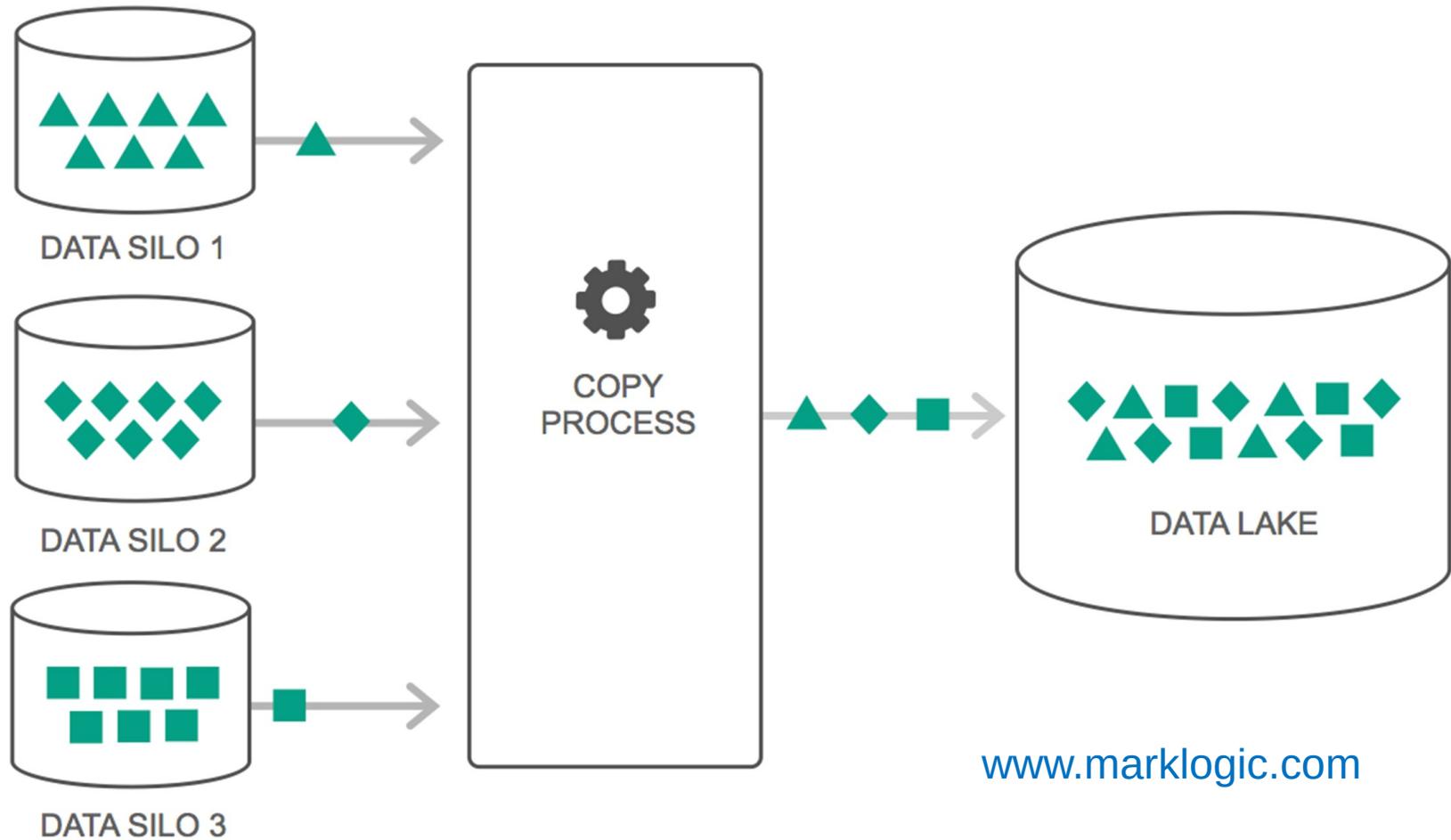


@luminousmen.com

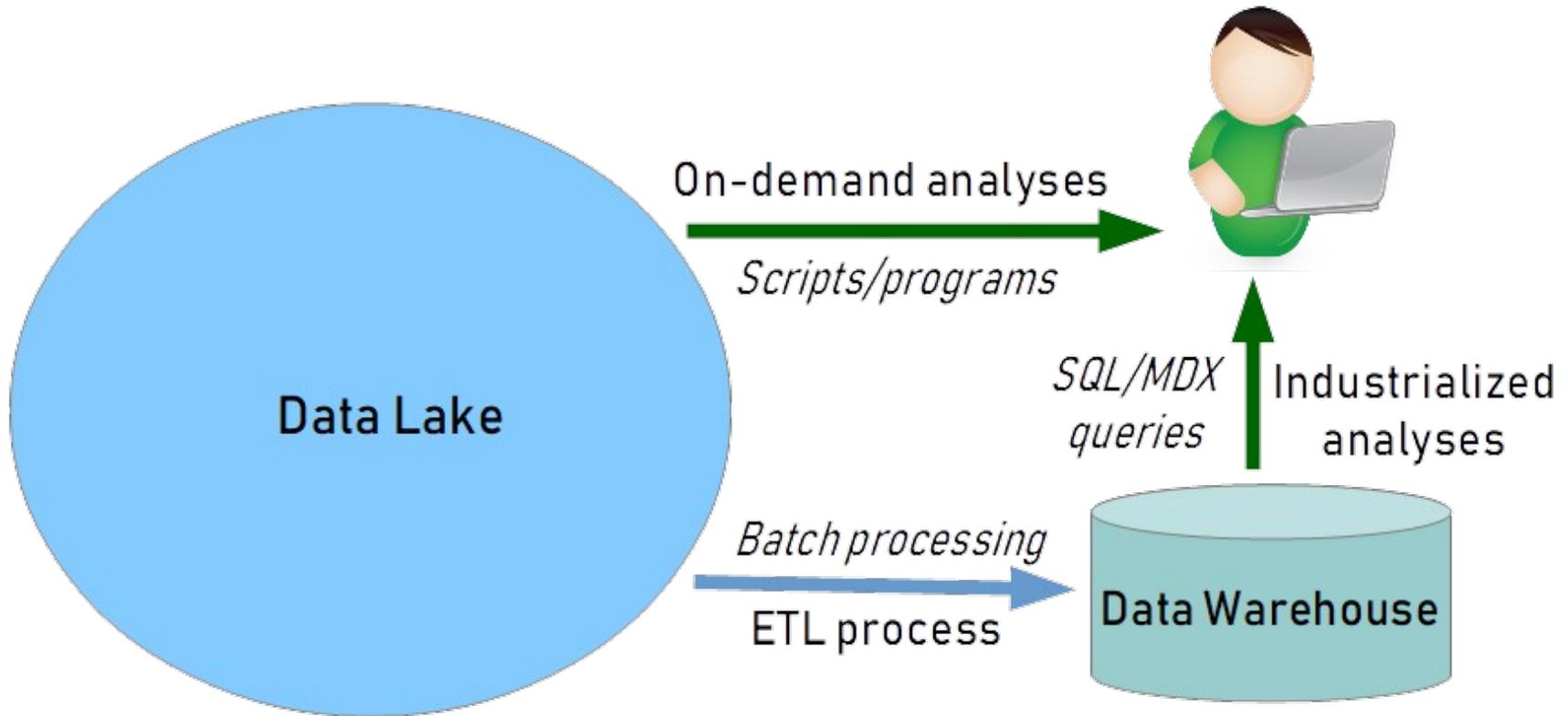
# Data lake vs. Data warehouse

Domaine	Entrepôt de données	Lac de données
Données	Nettoyées	Brutes
Schéma	<i>Schema on write</i>	<i>Schema on read</i>
Sources de données	En nombre limité	Multiples
Structure	Données structurées	Données structurées, semi-structurées, non-structurées
Analyses	Industrialisées	À la demande
Accès aux données	SQL/MDX	Scripts/programmes
Coût de stockage	Peu coûteux	Très peu coûteux
Intégration données	ETL	ELT
Perte d'information	Agrégation	Aucune
Architecture	Figée, silos	Flexible
Utilisateurs	Experts métier	<i>Data scientists</i>

# Entrepôt comme source d'un lac



# Lac comme source d'un entrepôt



Sawadogo et Darmont 2019 

# Entrepôt dans un lac

*(Raw data pond)*

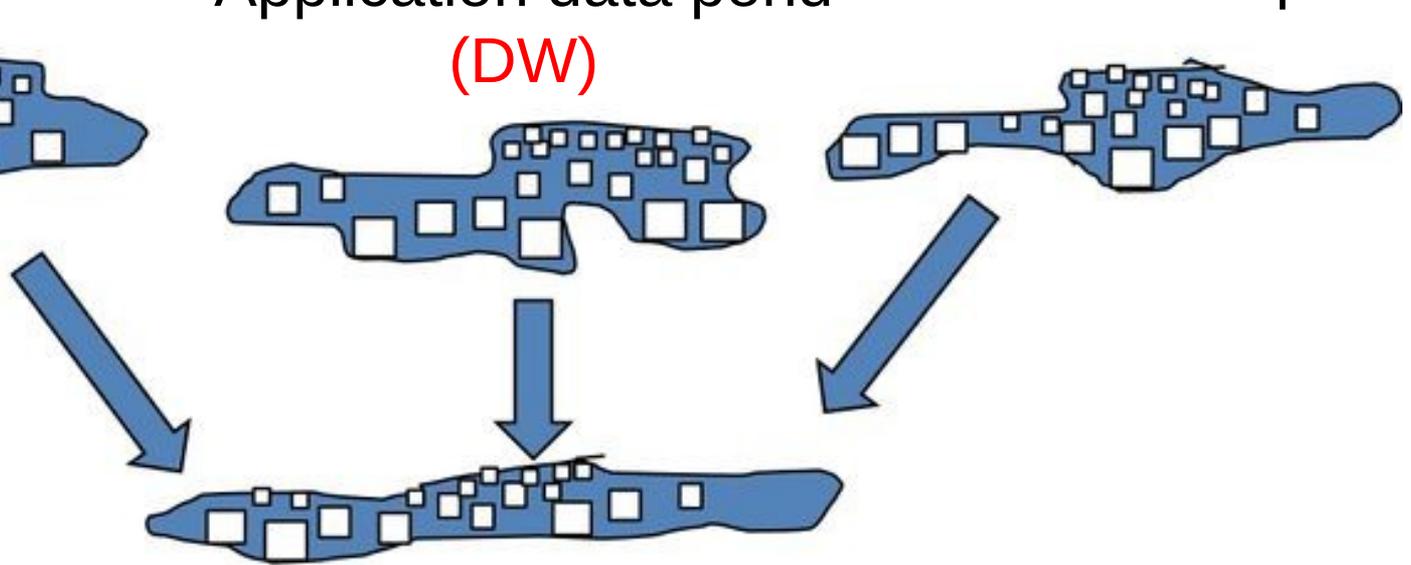
Analog data pond

Application data pond

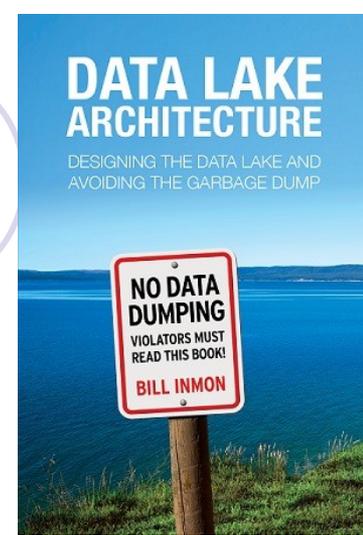
**(DW)**

Textual data pond

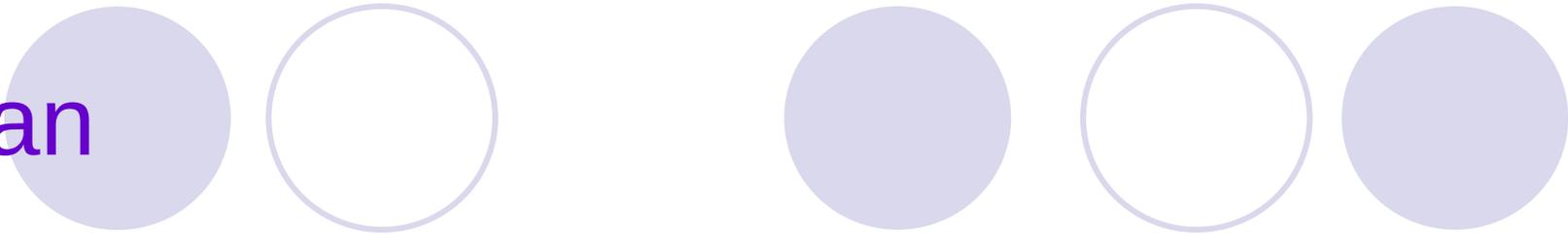
bassins  
de  
données



Archival data pond

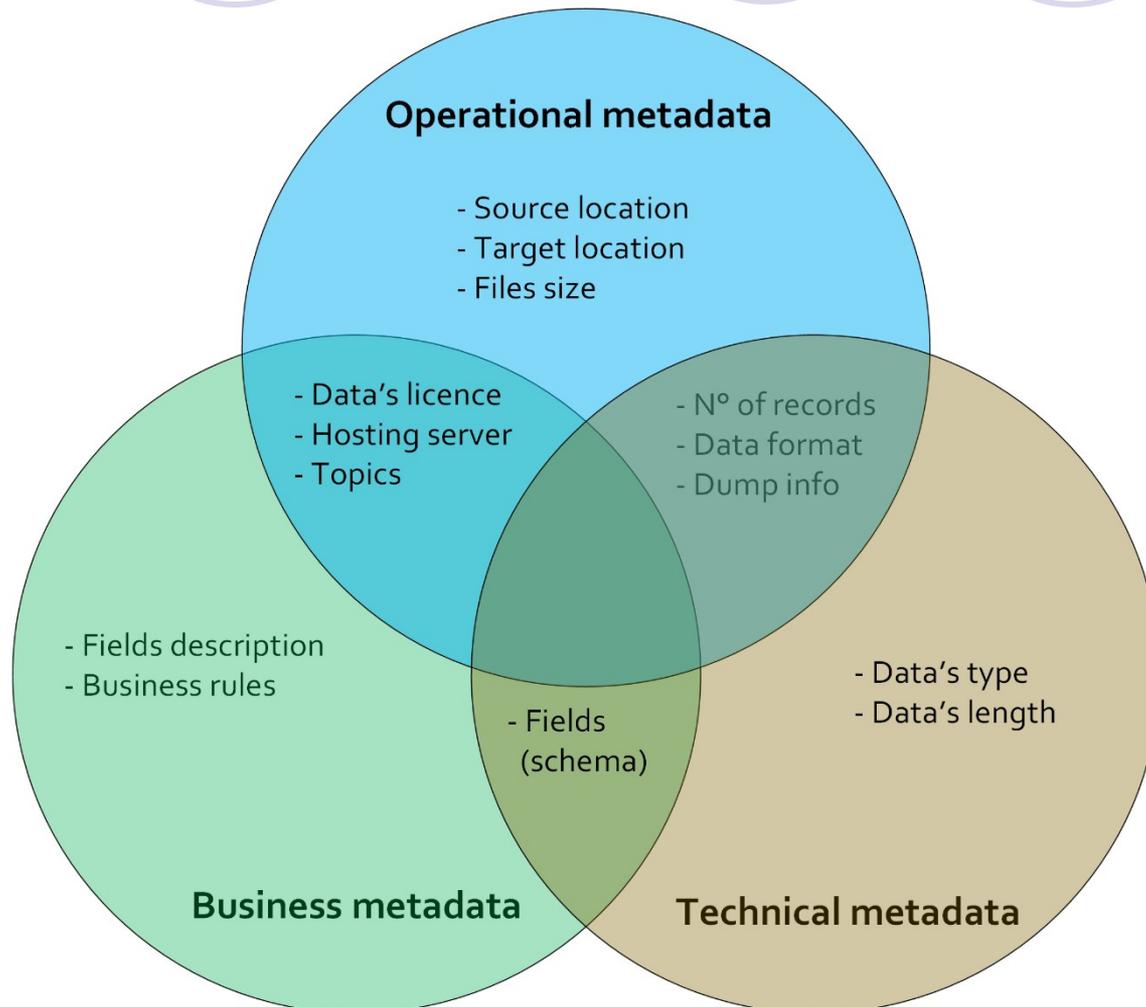


# Plan



- ✓ Définitions
- ✓ Entrepôts et lacs de données
- Métadonnées et modèles de métadonnées
- Architectures et technologies pour les lacs
- Discussion, travaux de recherche

# Typologie de Diamantini et al. (2018)



# Typologie de Sawadogo et Darmont (2019)

Généralisation de Maccioni and Torlone (2018)

**Objet** = ensemble de données homogènes  
Table relationnelle, fichiers (XML, tableur, texte, multimédia...)

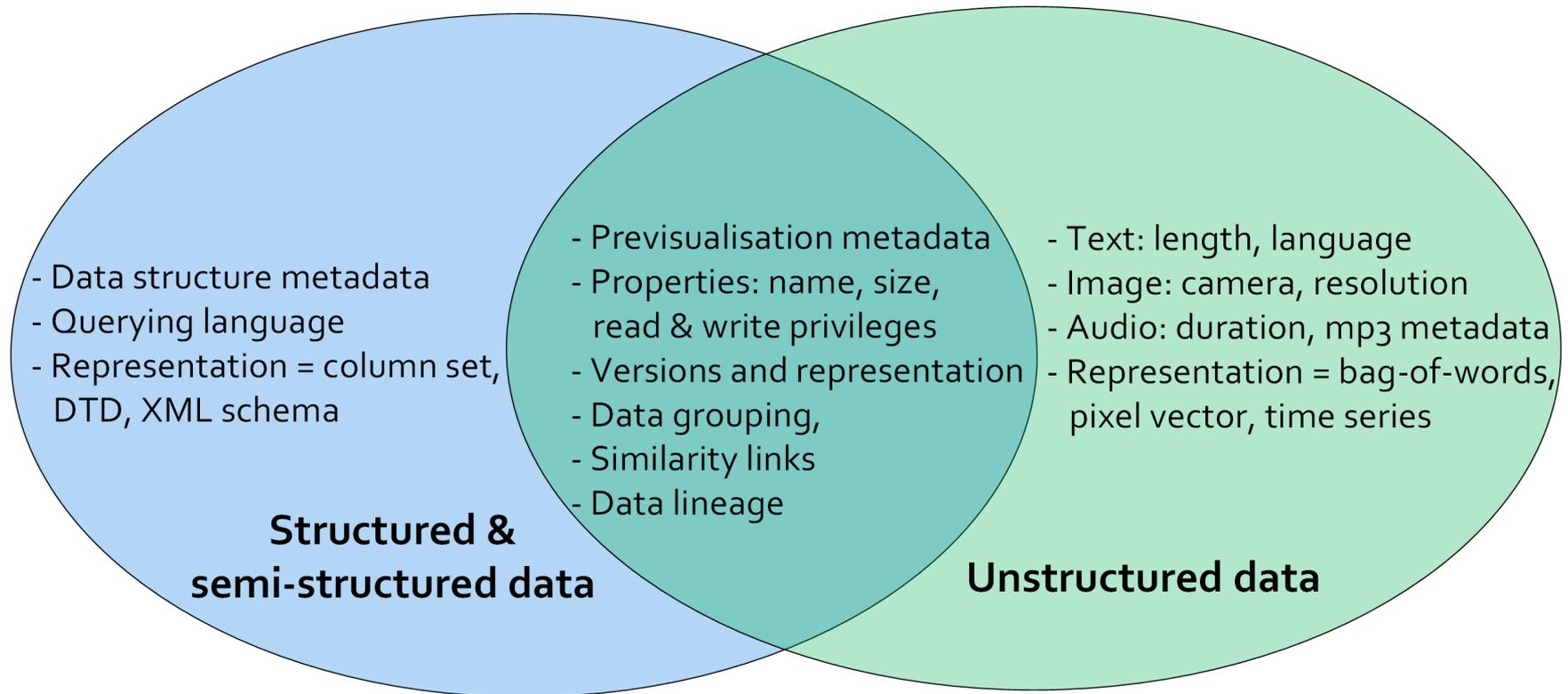
Métadonnées intra-objet	Exemples
Propriétés	Nom de fichier, taille, date de création...
Prévisualisations/résumés	Schéma, nuage de mots...
Versions et représentations	Transformation des données
Métadonnées sémantiques	Description, catégorie...

# Typologie de Sawadogo et Darmont (2019)

Métadonnées inter-objets	Exemples
Regroupements	Thématiques, par langue...
Similarités	Via des mesures de similarité
Parentés	Jointures, unions...

Métadonnées globales	Exemples
Ressources sémantiques	Ontologies, taxonomies...
Index	Index inversés
Journaux	<i>Logs</i>

# Typologie opérationnelle

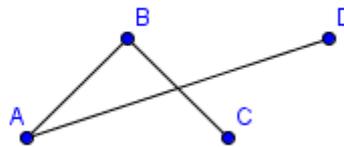
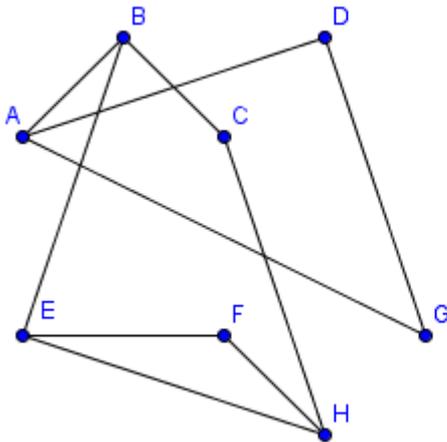


Sawadogo et Darmont 2019 

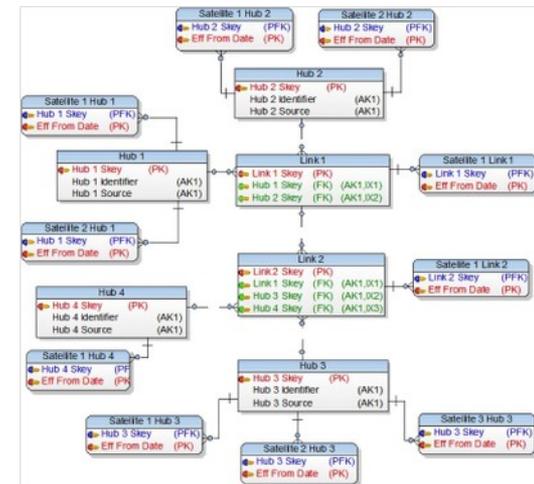
# Représentation logique des métadonnées

## Modèles relationnels ou NoSQL

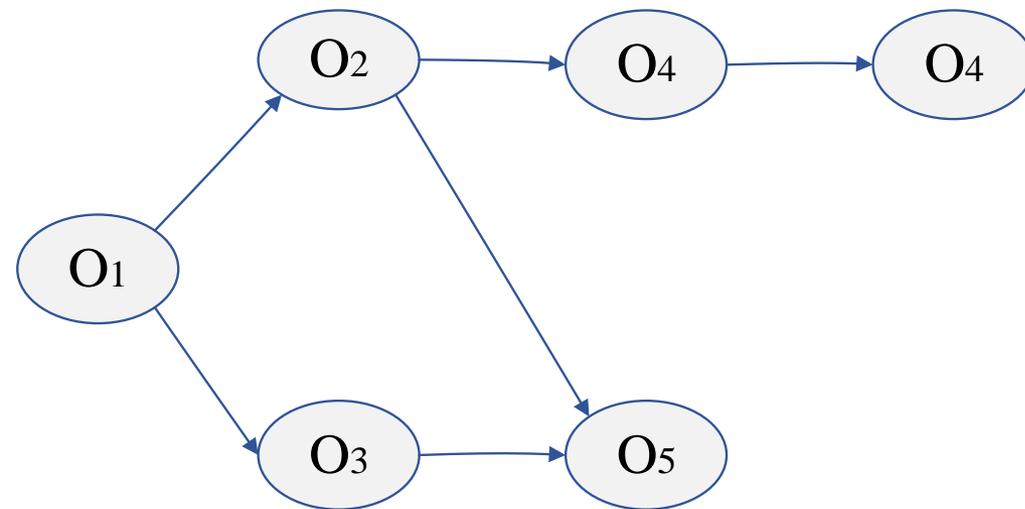
### Graphes



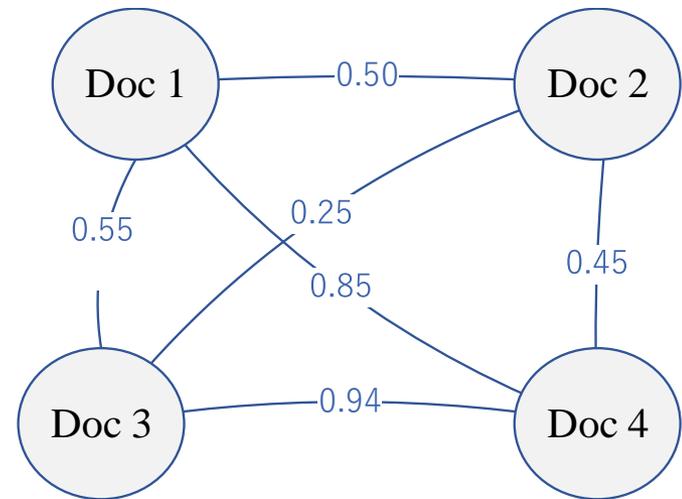
### Modélisation ensembliste



# Ex. de métadonnées en graphes



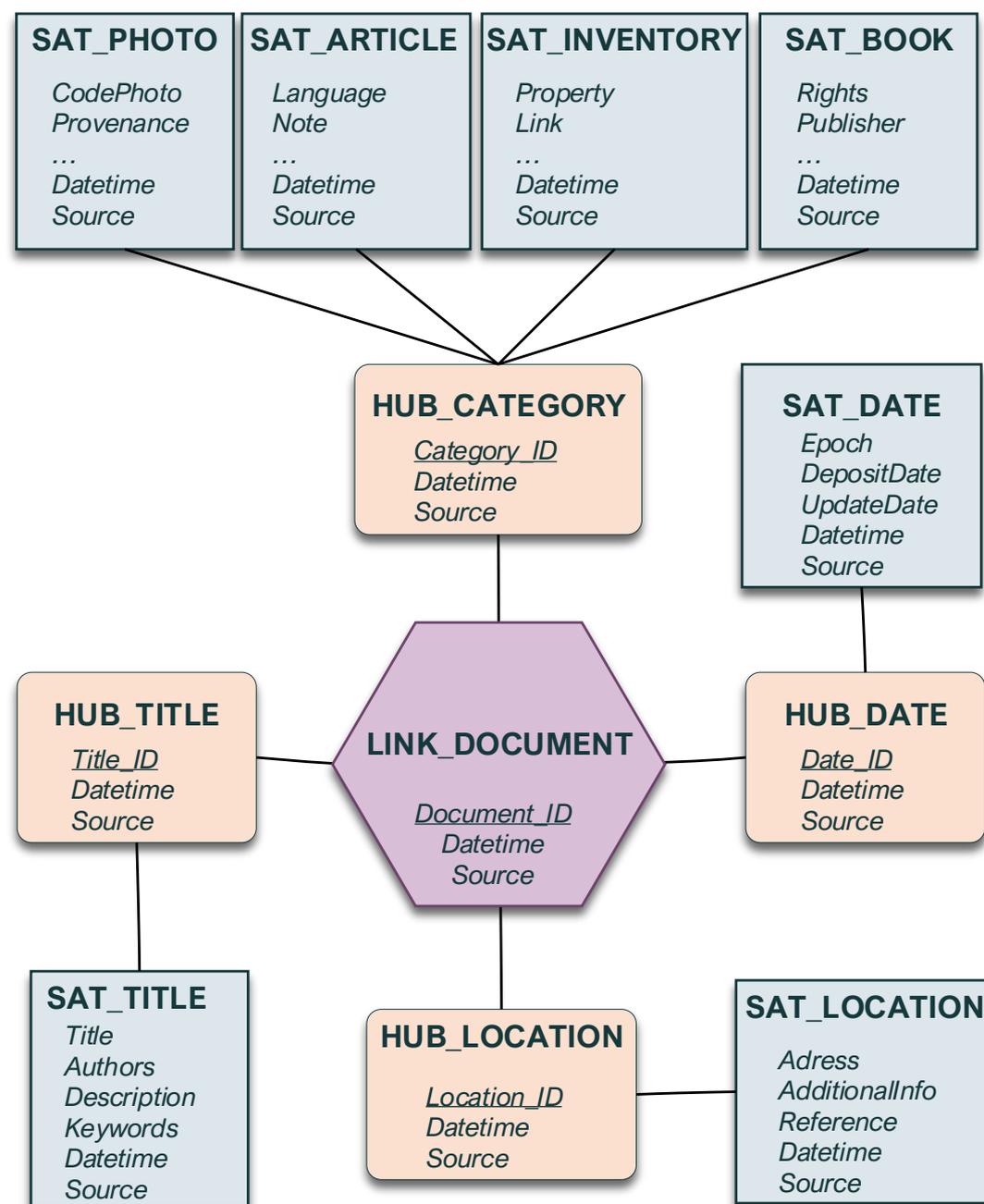
a) Provenance graph



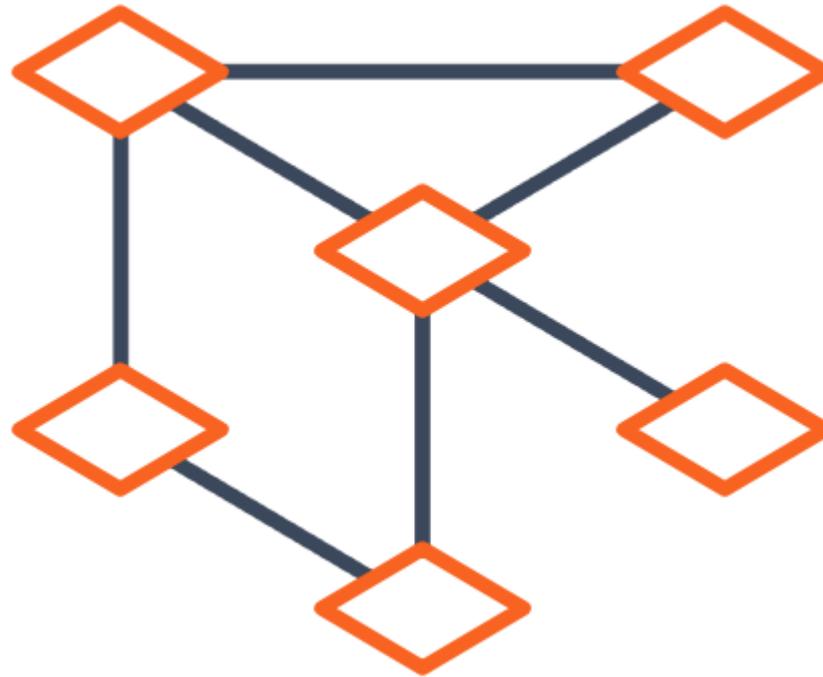
b) Similarity graph

# Ex. de métadonnées en data vault

Nogueira et al. 2018 

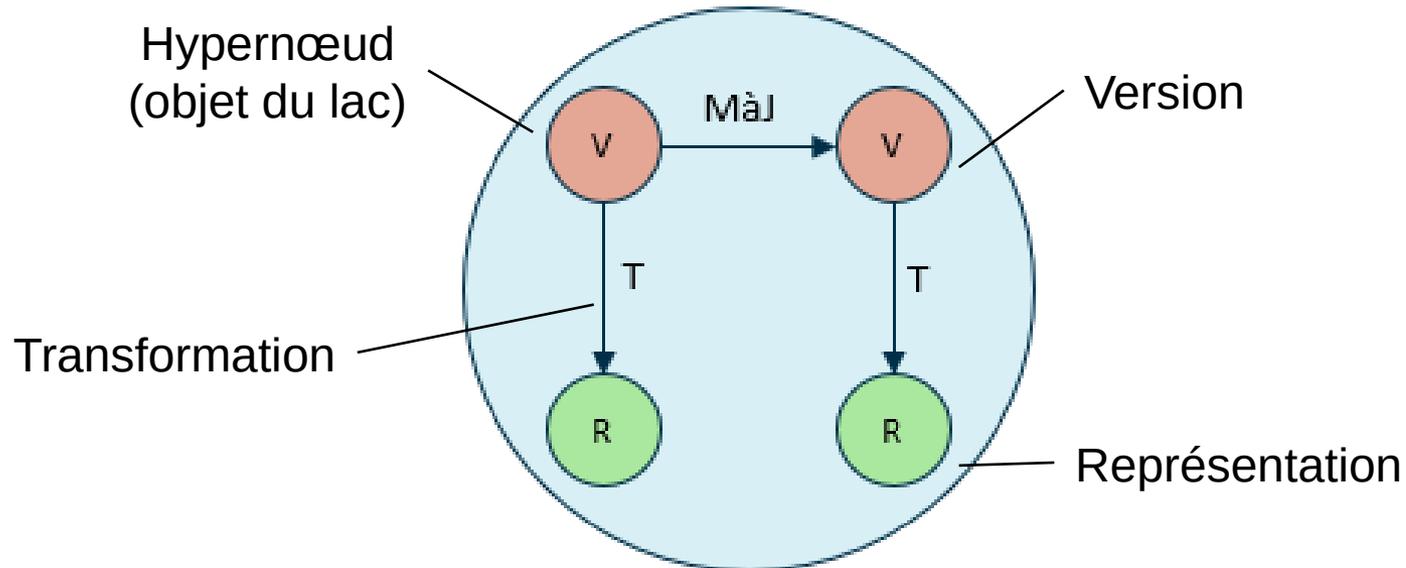


# Modèles de métadonnées récents



[blog.tensorflow.org](https://blog.tensorflow.org)

## Graphe d'hypernœuds

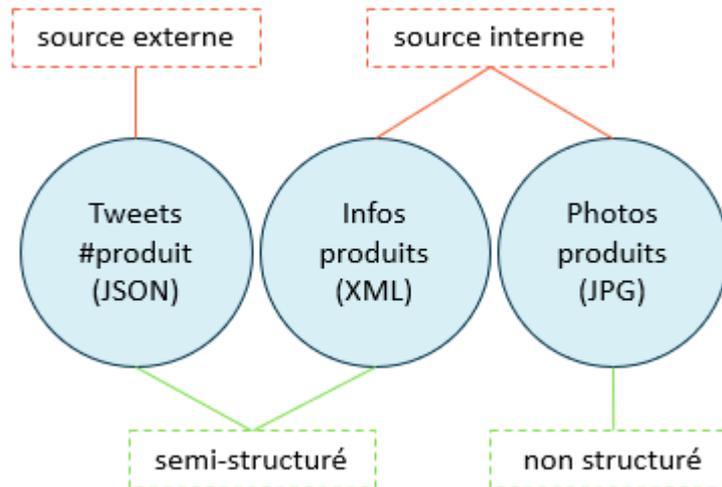


Nœuds et arcs portent des attributs (métadonnées intra-objet).

# MEDAL

- Métadonnées inter-objets

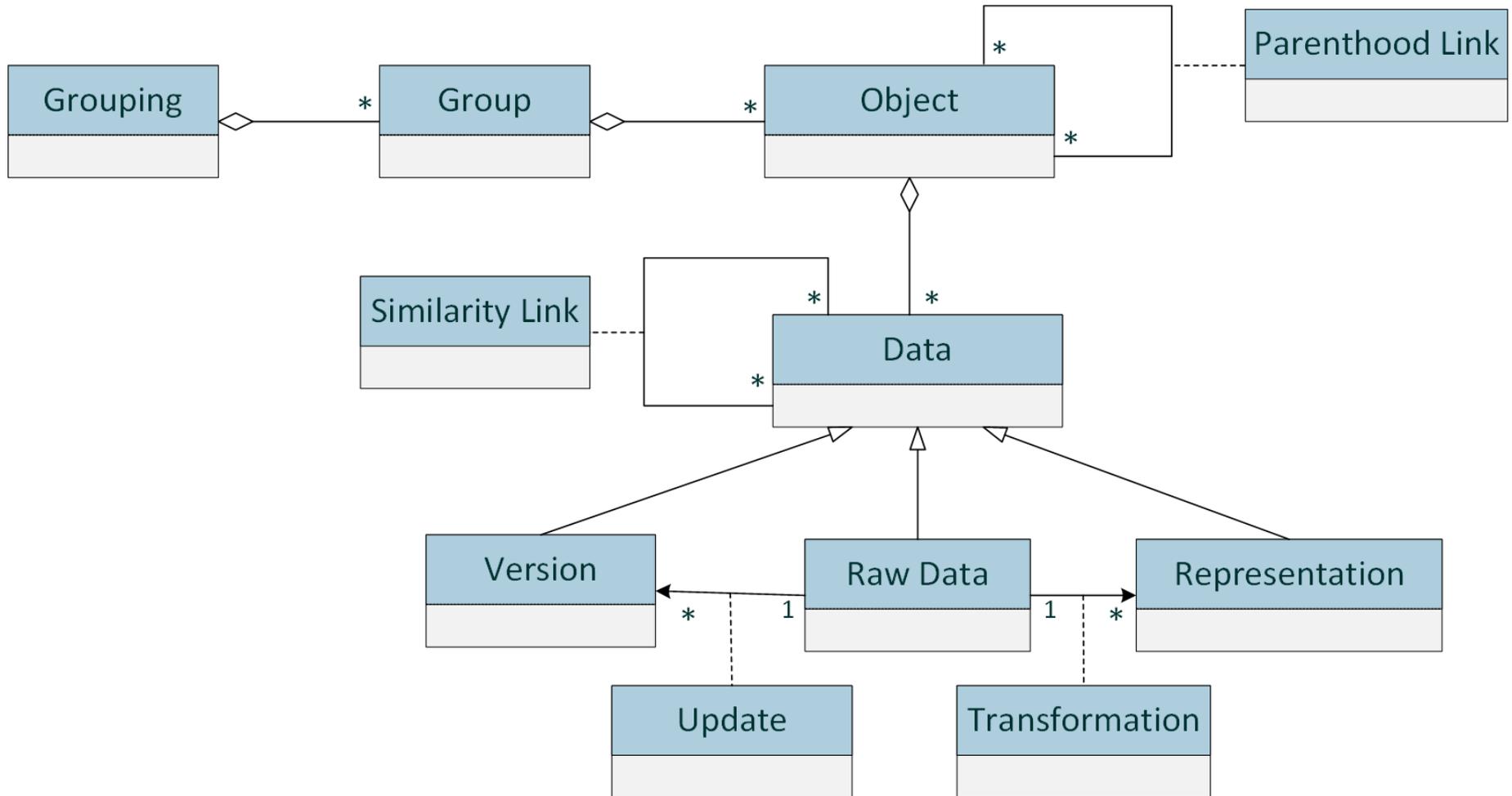
- Exemple de regroupements



- Métadonnées globales

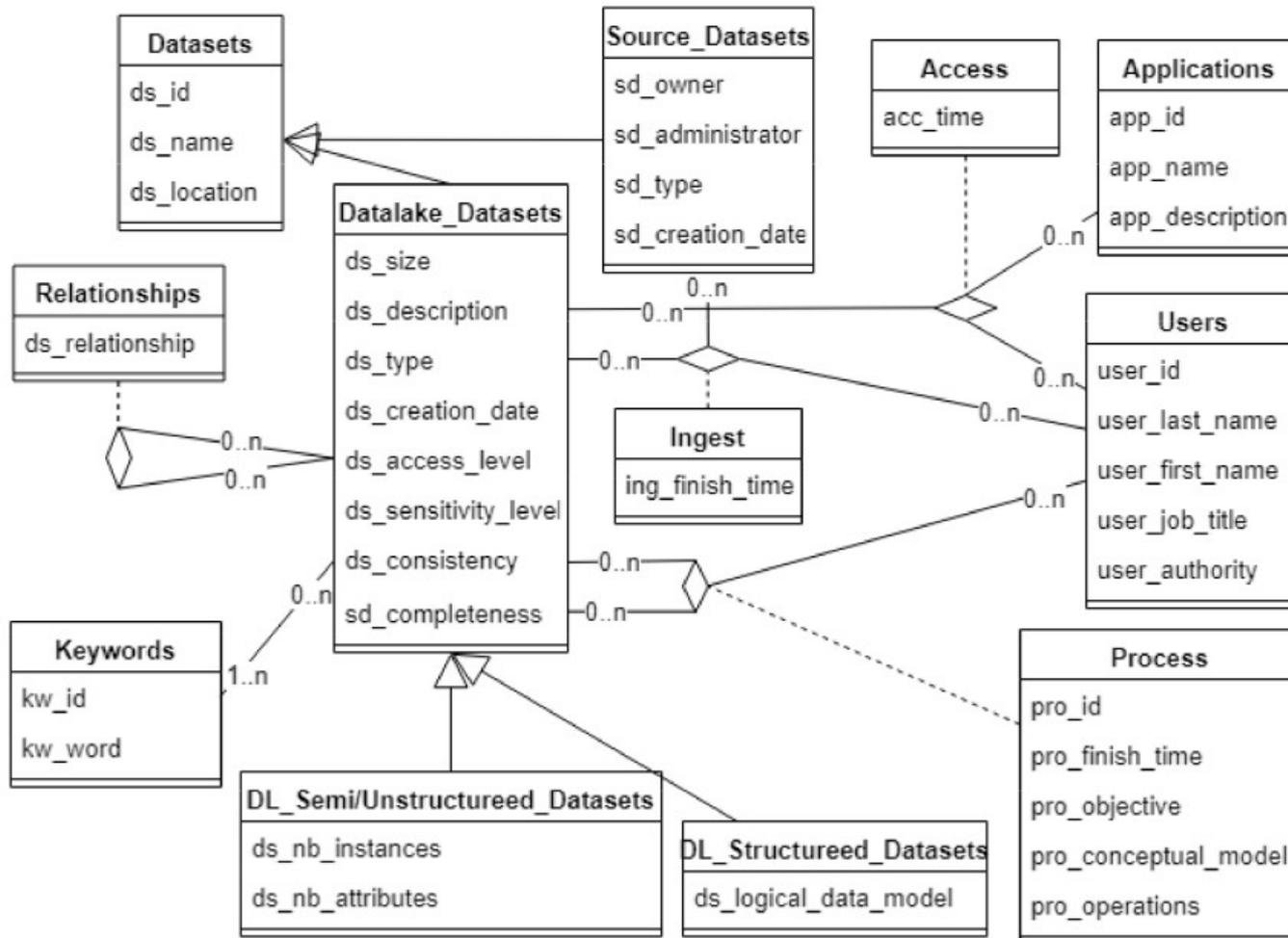
- Dans des nœuds non connectés au reste du graphe

# MEDAL : Modèle conceptuel



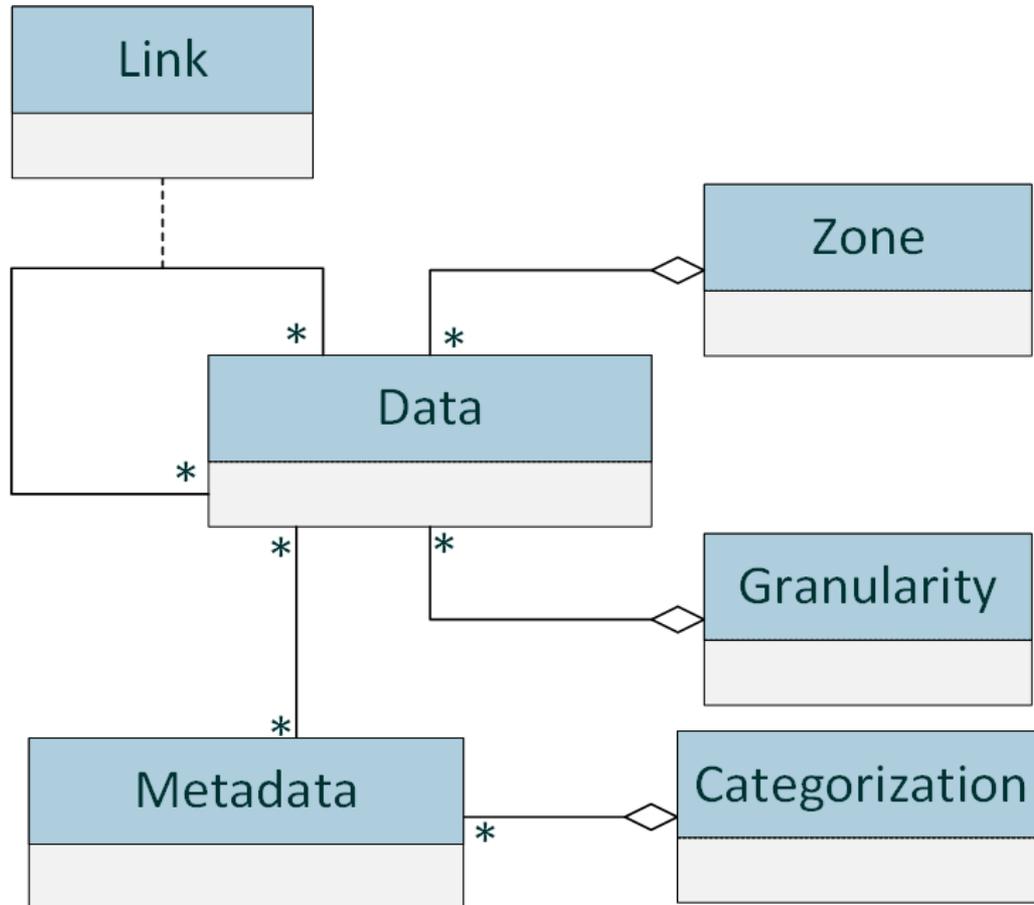


# DAMMS dans le détail



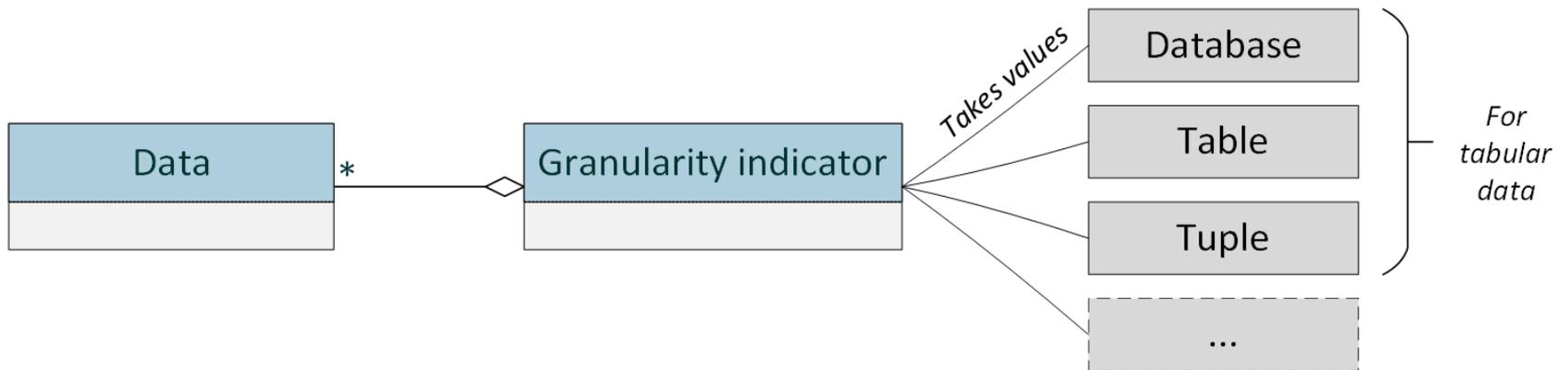
# HANDLE

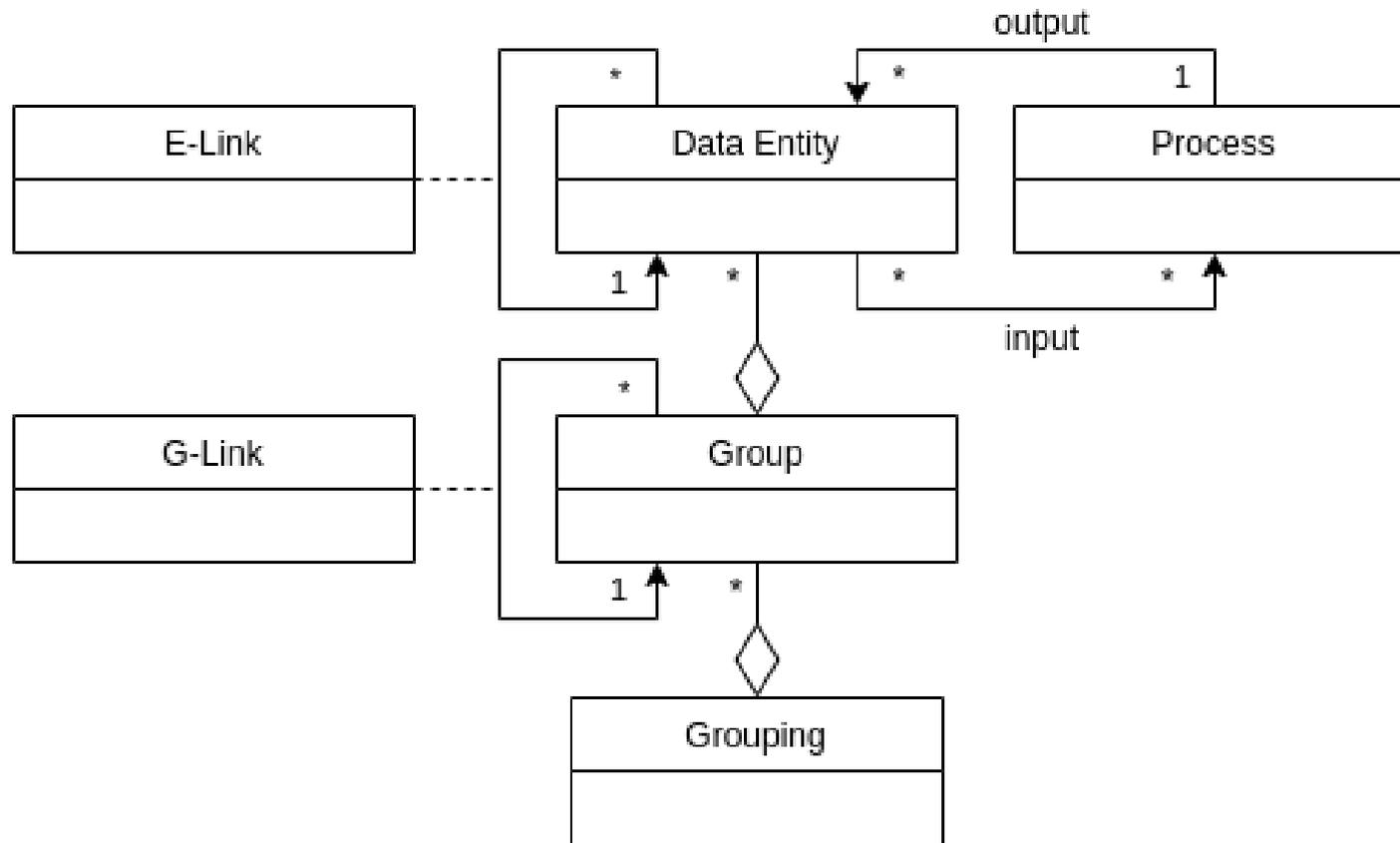
Eichler et al. 2020

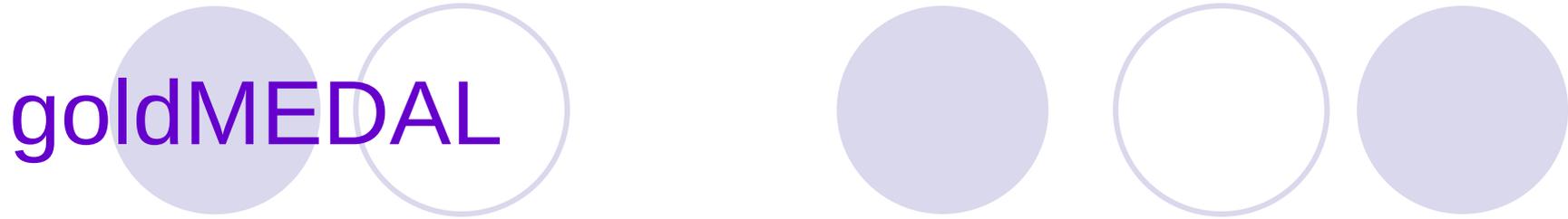


# Extensions HANDLE

Exemple : extension granularités (existent aussi pour les zones et catégorisations)







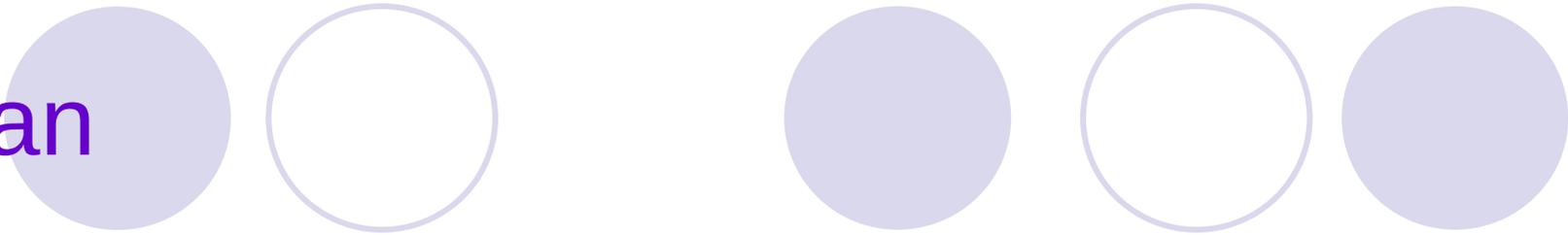
# goldMEDAL

- Généralise les concepts des autres modèles
  - Zones et granularité à l'aide des groupements
- Possibilité explicite de modéliser le lignage
  - Avec les processus
  - Dynamique des données et des métadonnées
- Modèles conceptuels, logique et physique
- **Le plus générique aujourd'hui !** (métamodèle)

# Comparaison des modèles de métadonnées

Caractéristiques	MEDAL	DAMMS	HANDLE	goldMEDAL
Enrichissement sémantique	X	X	X	X
Polymorphism/zones multiples	X	X	X	X
Versionnement des données	X	X		X
Suivi des usages	X	X	X	X
Catégorisation	X	X	X	X
Liens de similarité	X	X	X	X
Propriétés des métadonnées	X	X	X	X
Granularités multiples			X	X
<b>Total</b>	<b>7/8</b>	<b>7/8</b>	<b>7/8</b>	<b>8/8</b>

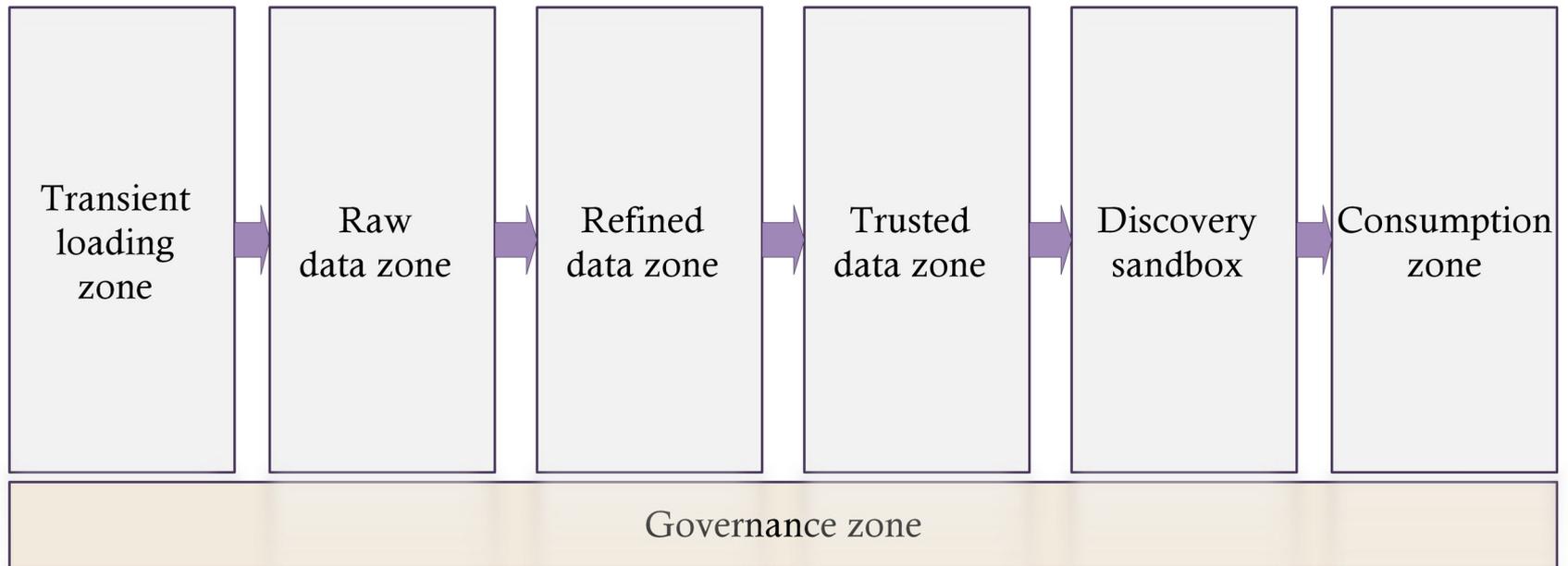
# Plan



- ✓ Définitions
- ✓ Entrepôts et lacs de données
- ✓ Métadonnées et modèles de métadonnées
- Architectures et technologies pour les lacs
- Discussion, travaux de recherche

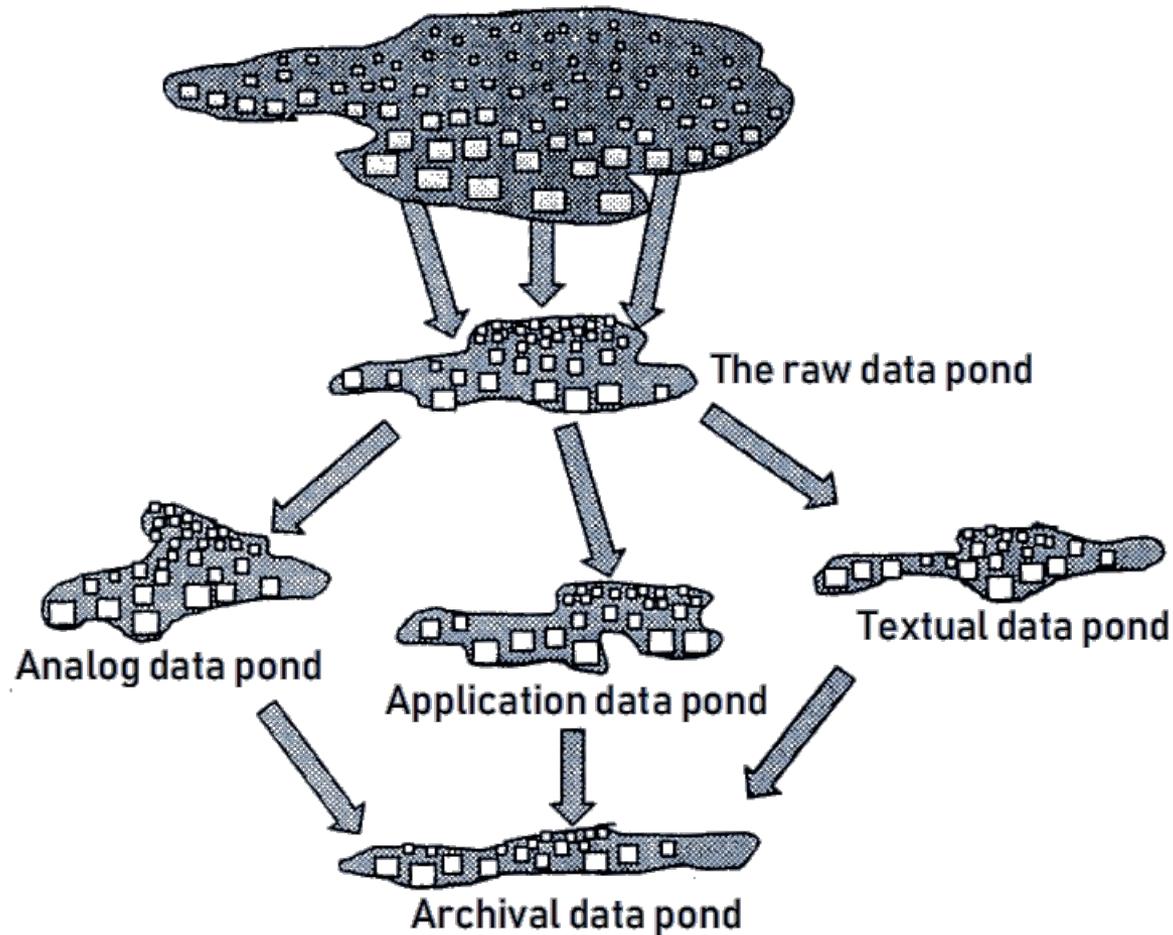
# Architecture en zones

Zaloni 2015



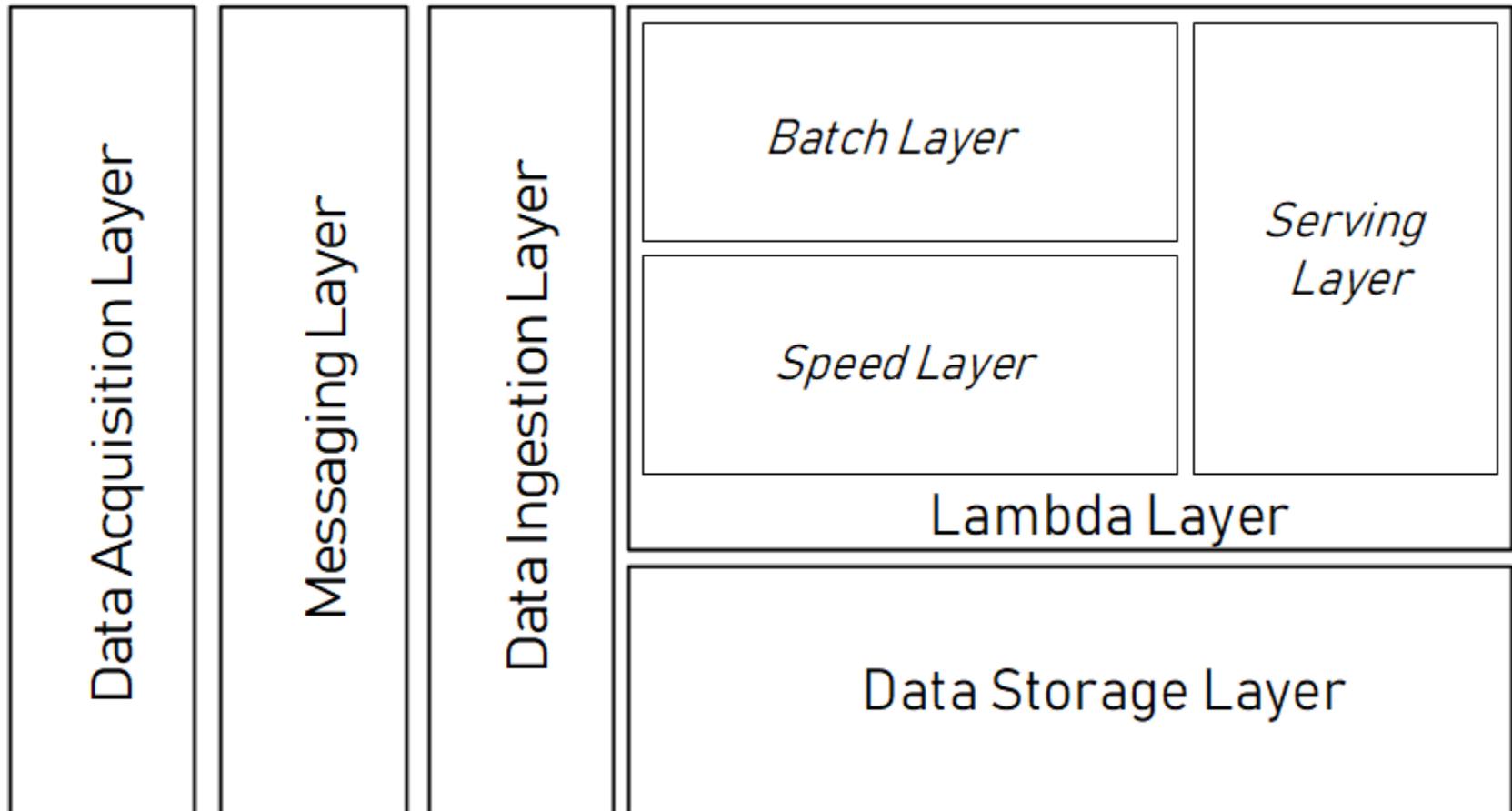
# Architecture en bassins

Inmon 2016



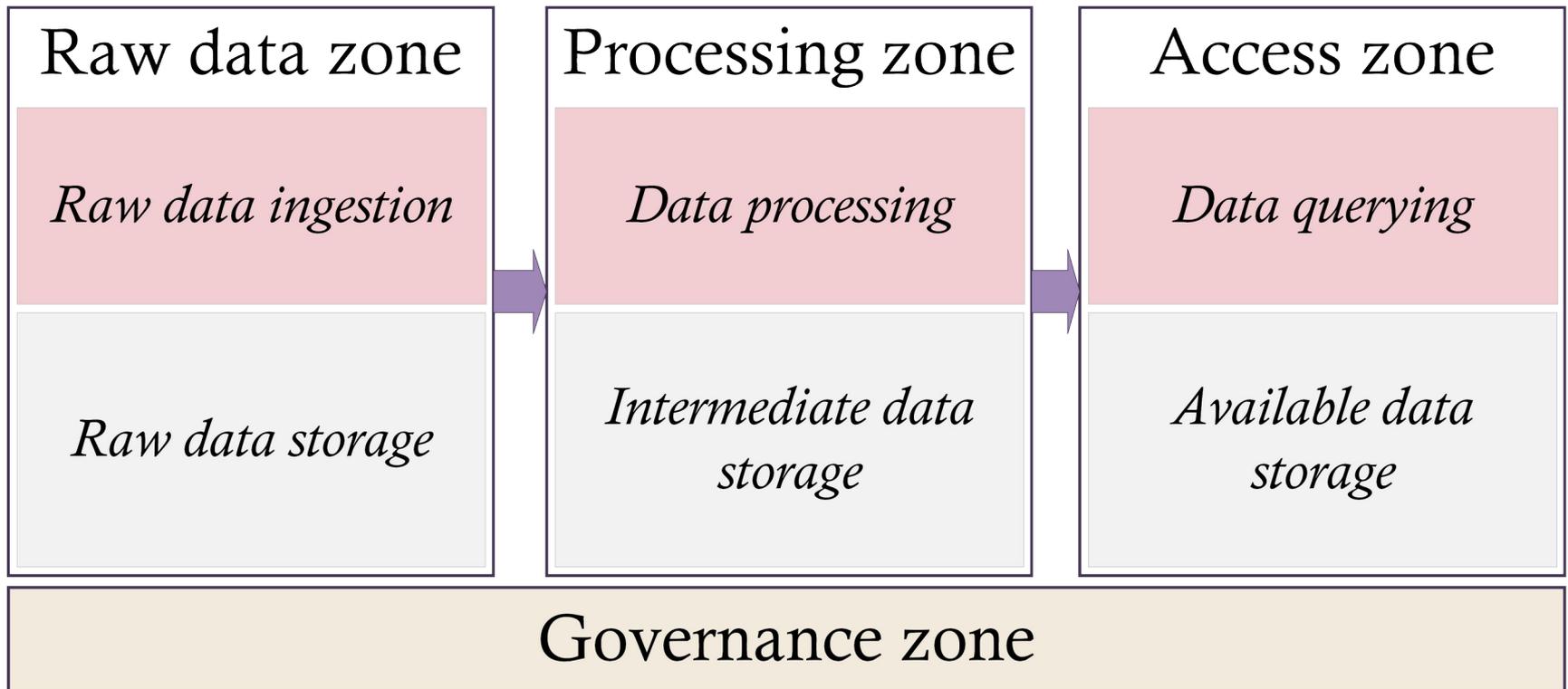
# Architecture lambda

John & Misra 2017

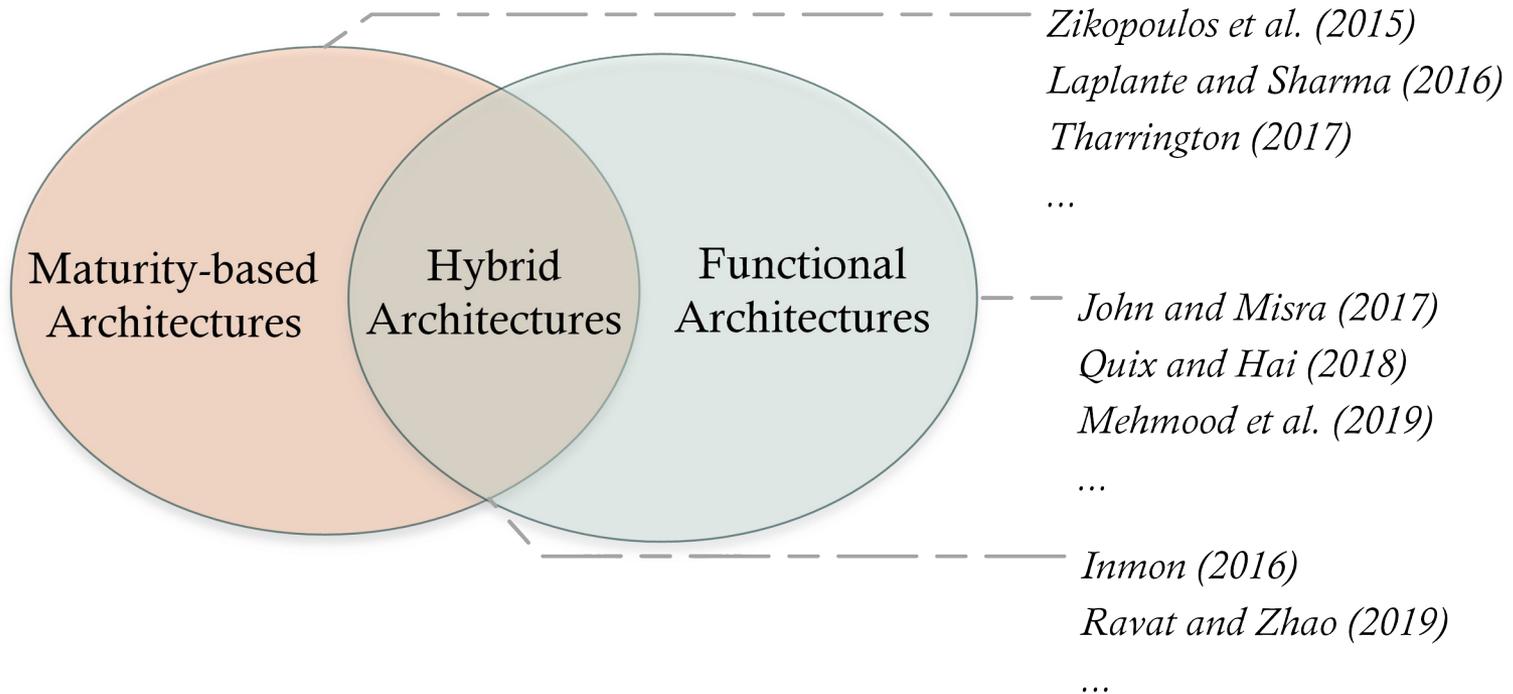


# Architecture hybride

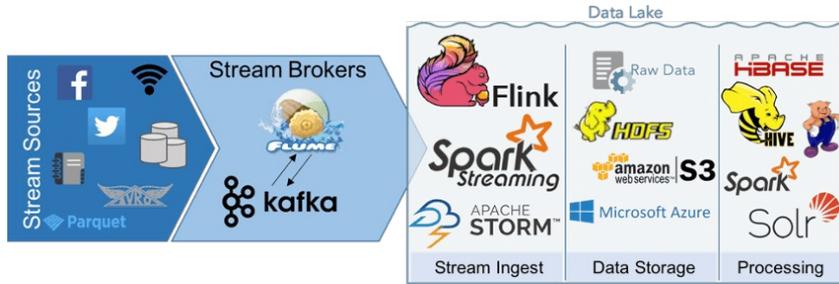
Ravat & Zhao 2019



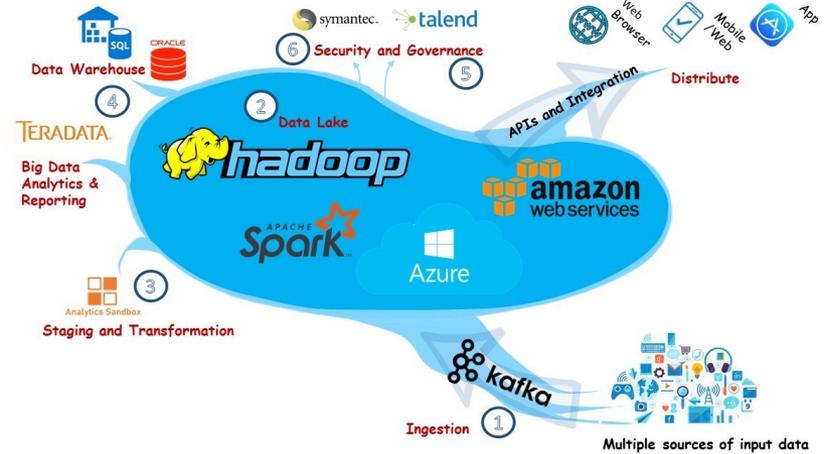
# Architectures fonctionnalité × maturité



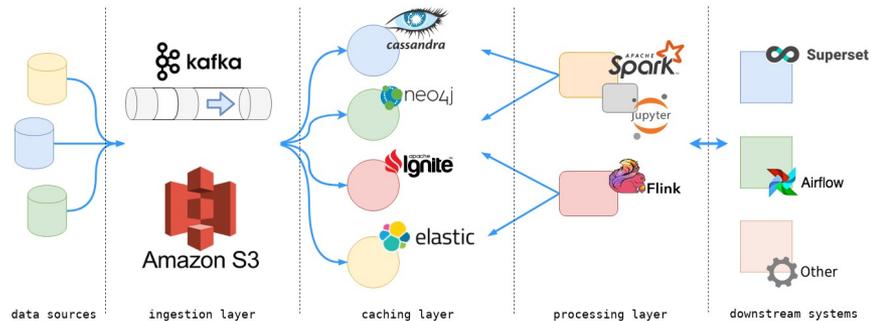
# Technos pour les lacs de données



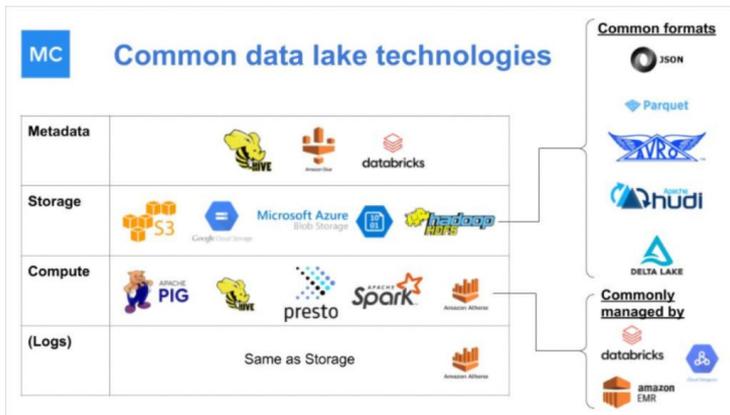
7wdata.be



kms-world.com

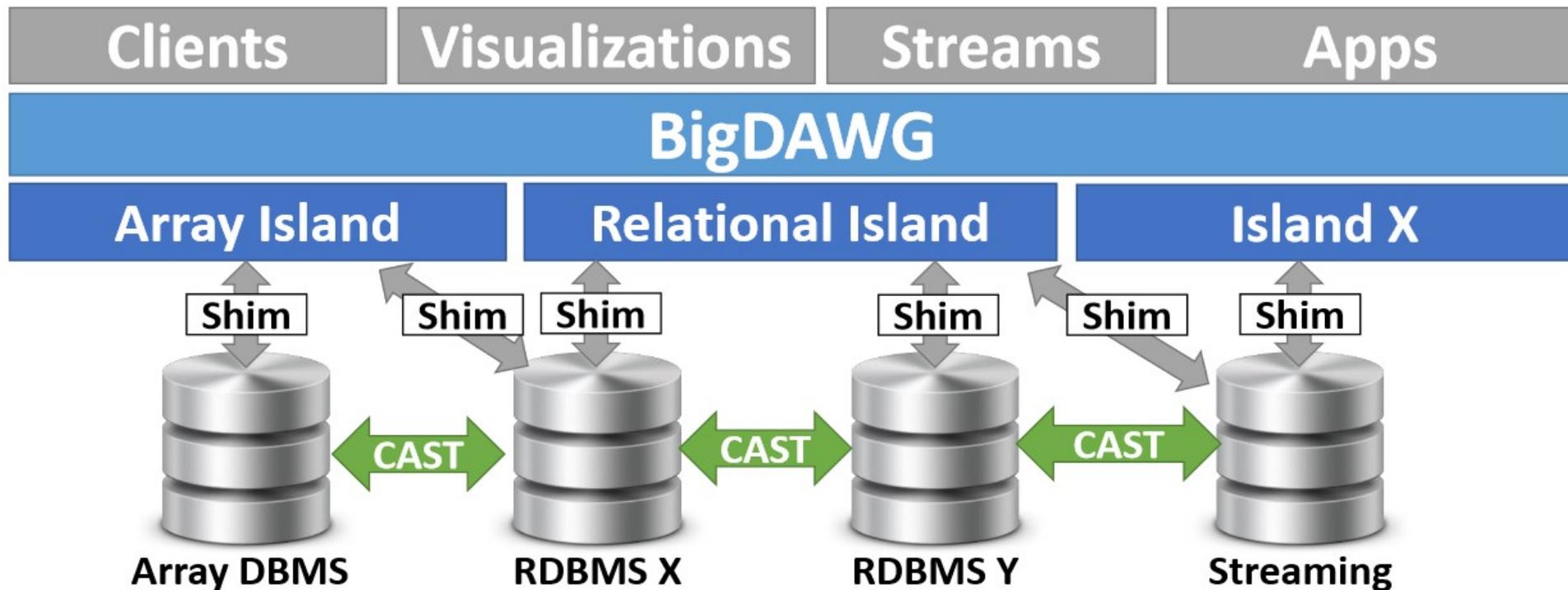


smartcat.io



www.montecarlo.com

# Polystores (ex. BigDAWG)



Jennie Duggan, Aaron J. Elmore, Michael Stonebraker, Magda Balazinska, Bill Howe, Jeremy Kepner, Sam Madden, David Maier, Tim Mattson, and Stan Zdonik. 2015. The BigDAWG Polystore System. SIGMOD Rec. 44, 2 (August 2015), 11-16. DOI=<http://dx.doi.org/10.1145/2814710.2814713>

# Ingestion et stockage des métadonnées



Ingestion des données



Métadonnées de base



# Apache Atlas

Extraction avancée et gestion des métadonnées

# Gestion des métadonnées

Tout SGBD relationnel ou NoSQL



mongoDB

Les spécialistes

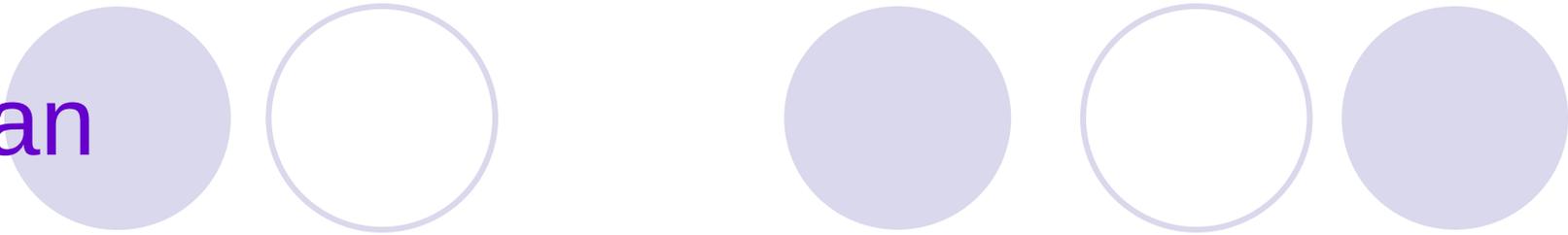


Apache Atlas



Open  
Metadata

# Plan



- ✓ Définitions
- ✓ Entrepôts et lacs de données
- ✓ Métadonnées et modèles de métadonnées
- ✓ Architectures et technologies pour les lacs
- Discussion, travaux de recherche

# Les lacs de données sur le grill



Agilité et flexibilité

Faible coût de stockage

Fidélité aux données

Gestion de données non structurées

Intégration des données en temps réel

Détection de relations entre données

Analyses à la volée

Passage à l'échelle

Tolérance aux pannes

Exploitation de données tierces

Incompatibilité avec des méthodes

Incohérences de données

Confusion autour du concept de lac

Besoin d'interfaces d'accès données

Besoin de métadonnées adaptées

Manque de standards de gouvernance

Compétences pour la mise en œuvre

Sécurité en maturation

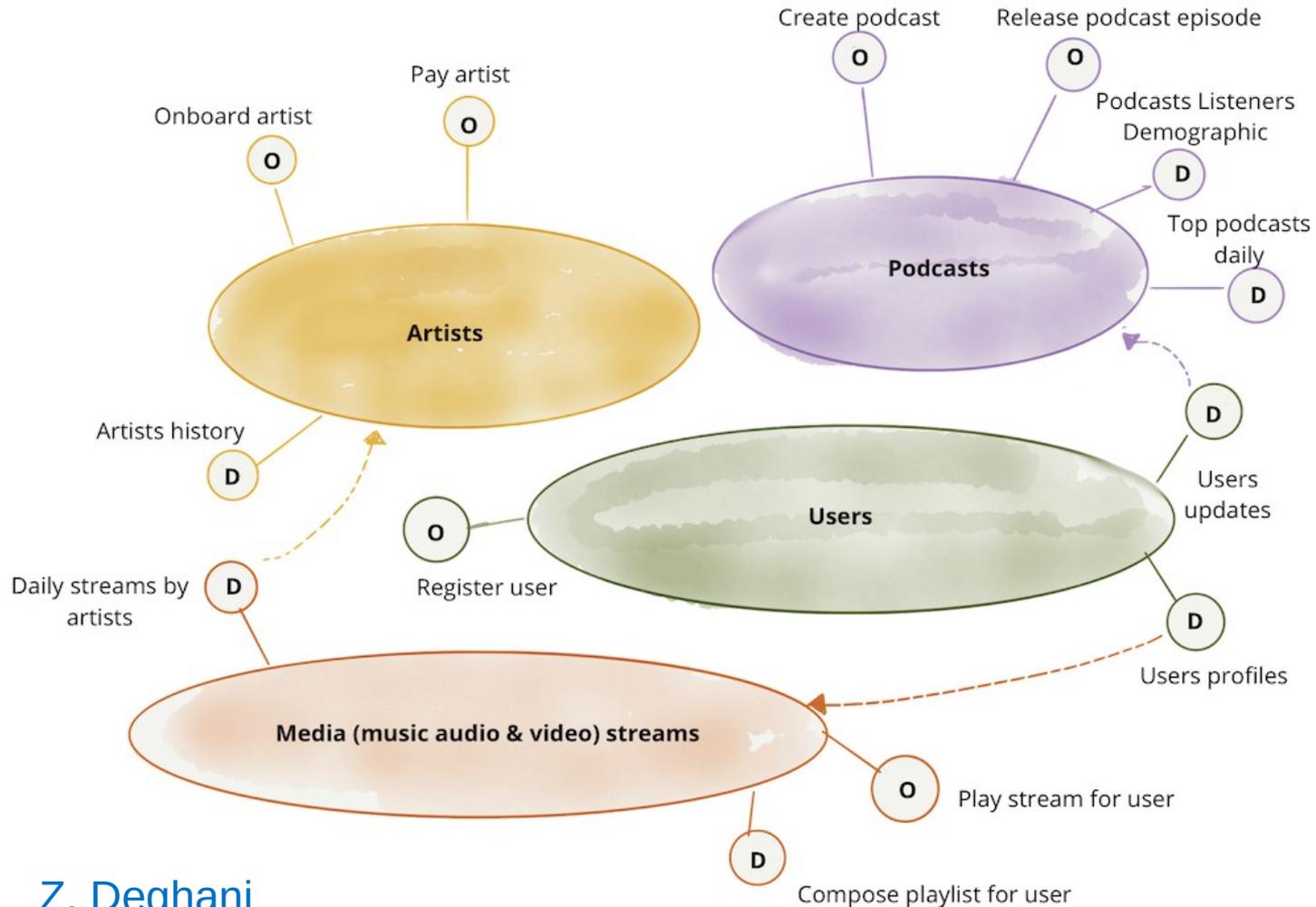
**Monolithiques, « challengés » par les maillages de données (*data meshes*)**

# Data mesh ou la décentralisation

Zhamak Dehghani 2019 (consultante principale, Thoughtworks)

- Objectif : passage à l'échelle
  - Changement rapide des données et de leurs modèles
  - Croissance permanente des producteurs de données
  - Augmentation des consommateurs de données (ML)
- Les 4 piliers du data mesh
  - Découpage en domaines de données
  - Gestion des données en tant que produits
  - Infrastructure « self-service »
  - Gouvernance fédérée des données

# Domaines : producteurs et consommateurs de données



# Problèmes de recherches actuels

- Intégration/transformation des données
  - Optimisation des *User-Defined Functions* (UDFs)
    - Ex. Tâches MapReduce
  - Ingestion en temps réel de données à haute vélocité
- Interrogation des données
  - Interopérabilité entre données (semi-)structurées et non-structurées
  - Gravité des données, intégration virtuelle
  - Performance

# Problèmes de recherches actuels

- Données non structurées
  - Génération de métadonnées
  - Données multimédia
- Gouvernance des données
  - Risques vs. problèmes à régler
  - Transformer les principes en solutions effectives
- Confidentialité des données
  - GDPR, quelqu'un ?

# Problèmes de recherches actuels

- « Industrialisation » des lacs de données
  - Pour des utilisateurs non spécialistes
  - Couche logicielle intermédiaire *et plus !*
- Données
  - Faciles à trouver, **A**ccessibles, **I**nteropérables, **R**éutilisables (principes FAIR)
- Maintenance des métadonnées
  - Évolution de catégories, de volume, de technos...

# Projets de recherche



Projet TECTONIQ  
(2015)



Stage 2015  
N. Pathinara



TER 2017  
I. Nogueira et  
M. Romdane



Projet COREL  
(2017-2018)



Stages 2018  
P. Sawadogo  
et T. Kibata



Projet AURA-PMI  
(2017-2021)



Thèse 2018-2021  
P. Sawadogo



Projet STRATEGE  
(2019)



Stage 2019  
R. Dib



Projet HyperThesau  
(2018-2019)



Postdoc P. Liu  
2019-2020



Projet LIFRANUM  
(2020-2023)



Postdoc  
J. Espinosa  
2020-2021



Stage  
V. Renault  
2022

# Projets de recherche



Projet DataLAC  
(2020-2023)



Stage  
L. Ciuraneta  
2021



Thèse CIFRE  
2022-2025  
A. Diouane



Projet  
PicassoLetters  
(2021-2022)



Stage  
I. Slalmi  
2022



Stage  
R. El-Idrissi  
2023



Thèse CDU  
R. El-Idrissi  
2023-2026



Stage  
A. Derder  
2023



Stage  
R. Aoudj  
2023



Stage  
M. Aliouche  
2024



Stage  
M. Sahrane  
2024



Stage  
A. Fugier  
2024



Stage  
E. Yetkin  
2024



Stage  
T. Lecomte  
2024

# Réseau de recherche *datalakers*

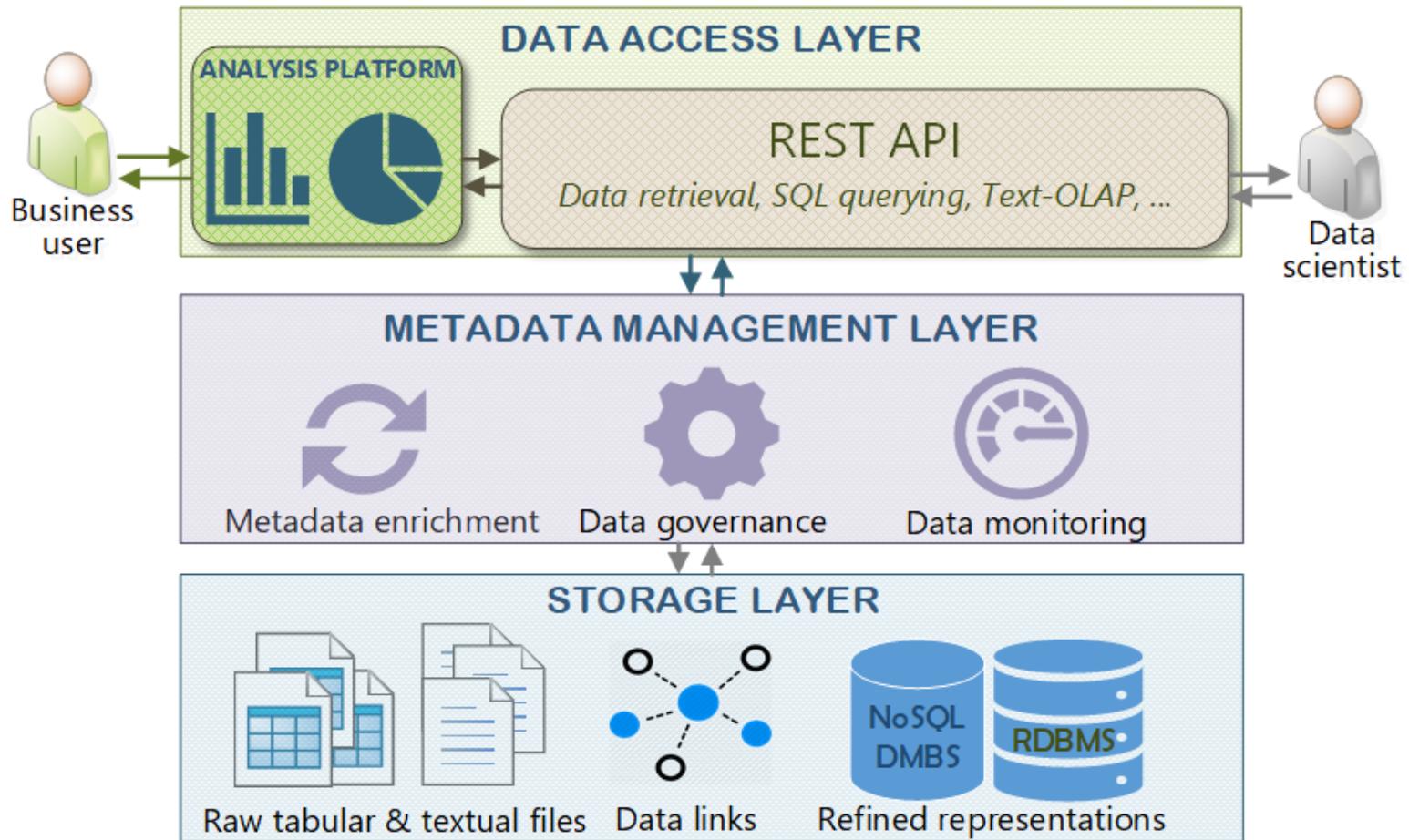


BIAL-X

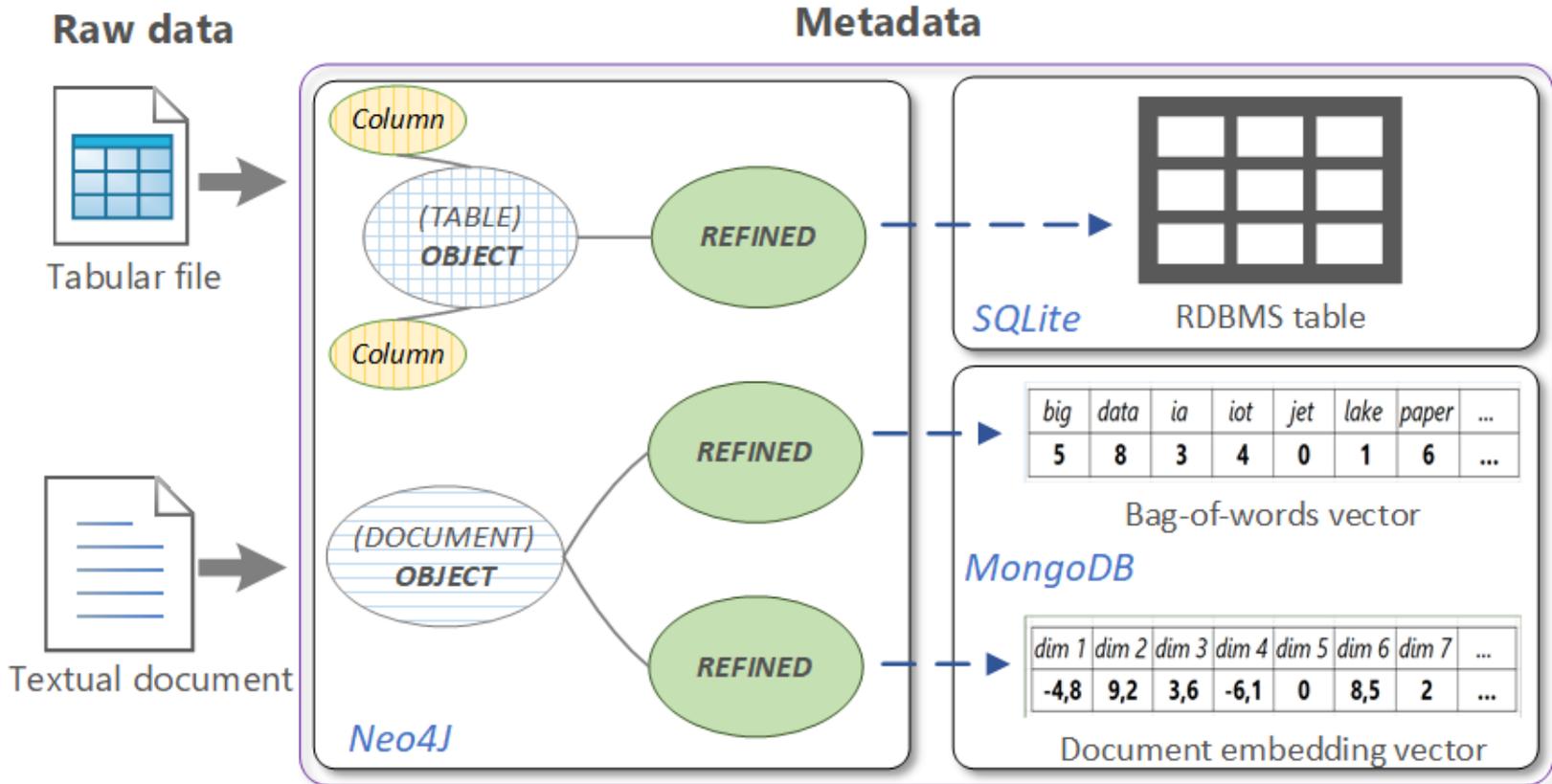


# Exemple : AUDAL

Sawadogo et Darmont 2021

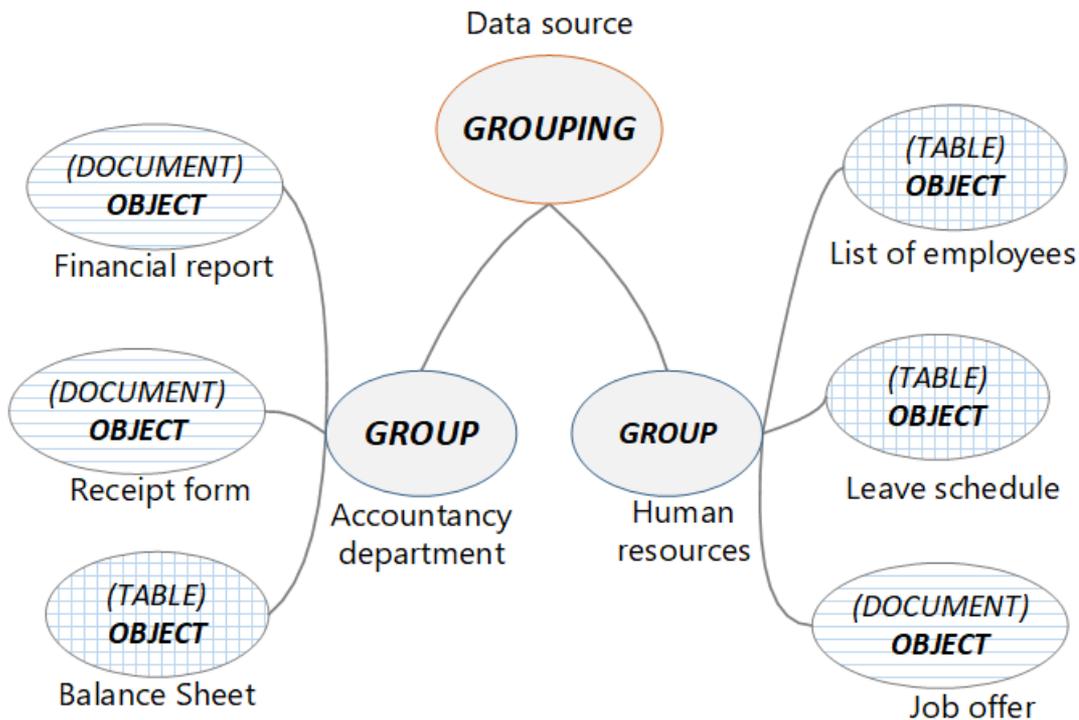


# Métadonnées intra-objet d'AUDAL

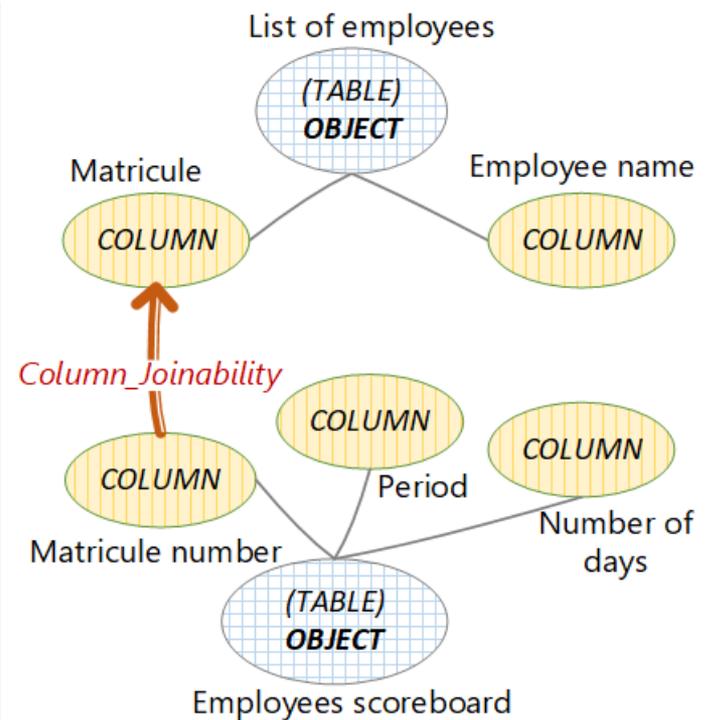


# Métadonnées inter-objets d'AUDAL

(A) Instance of grouping



(B) Instance of column joinability link



AUDAL Analysis Interface (1) (2) (3) alpha

### EXPLORATORY TASKS

Terms filtering +

Terms

+ matching ... +

- matching ... +

Parameters

Strictness **Any** All

Fuzzy search **Yes** No

Terms extension - None v +

Query Reset

Groupings -

Groups

- 1- cible
  - ALL
  - B2B [5804]
  - B2B-B2B2C [10]
  - B2B-B2C [1011]
  - B2C [1295]
- 2- digitalNativity
- 3- docCategory

### DOCUMENT PROPERTIES

#	title	
1	AG_11052020_DELFINGEN.pdf	👁
2	AG_13092018_AGM_SPINEWAY.pdf	👁
3	AG_16082019_SPINEWAY.pdf	👁
4	AG_25062018_AGM_SPINEWAY.pdf	👁
5	AG_26052020_SPINEWAY.pdf	👁
6	AG_28062019_AGM_SPINEWAY.pdf	👁
7	AG_28062019_SPINEWAY .pdf	👁
8	AG_28062019_SPINEWAY.pdf	👁
9	AGA_24962019_INTRASENSE.pdf	👁
10	AGE_2019_ARCHOS.pdf	👁
11	AGM 0 et E_05062020_DELFINGEN.pdf	👁

### ANALYSES

Documents Tables

Document properties ★

Parameters

Properties title [STRING v]

Visualisation Table v

Results

Agg. time (s)	0.366
Exp. time (s)	0.02
Result count	8120

Correlation analyses

Top Keywords

Highlights

Scoring

Links Analysis

Clustering

EXPLORATORY TASKS

DOCUMENT PROPERTIES

ANALYSES

Terms filtering +

Terms

+ matching ... +

- matching ... +

Parameters

Strictness **Any** All

Fuzzy search **Yes** No

Terms extension - None +

Query Reset

#	(1)	title	
1		AG_11052020_DELFINGEN.pdf	👁
2		AG_13092018_AGM_SPINEWAY.pdf	👁
3		AG_16082019_SPINEWAY.pdf	👁
4		AG_25062018_AGM_SPINEWAY.pdf	👁
5		AG_26052020_SPINEWAY.pdf	👁
6		AG_28062019_AGM_SPINEWAY.pdf	👁
7	(2)	AG_28062019_SPINEWAY .pdf	👁
8		AG_28062019_SPINEWAY.pdf	👁
9		AGA_24962019_INTRASENSE.pdf	👁
10		AGE_2019_ARCHOS.pdf	👁
11		AGM 0 et E_05062020_DELFINGEN.pdf	👁

Groupings -

Groups

- 1- cible
  - ALL
  - B2B [5804]
  - B2B-B2B2C [10]
  - B2B-B2C [1011]
  - B2C [1295]
- 2- digitalNativity
- 3- docCategorv

Documents Tables

Document properties

Parameters

Properties title [STRING]

Visualisation Table

Results

Agg. time (s)	0.366
Exp. time (s)	0.02
Result count	8120

Correlation analyses

Top Keywords

Highlights

Scoring

Links Analysis

Clustering

EXPLORATORY TASKS

Terms filtering

Terms

+ matching  +

- matching  +

Parameters

Strictness  Any  All

Fuzzy search  Yes  No

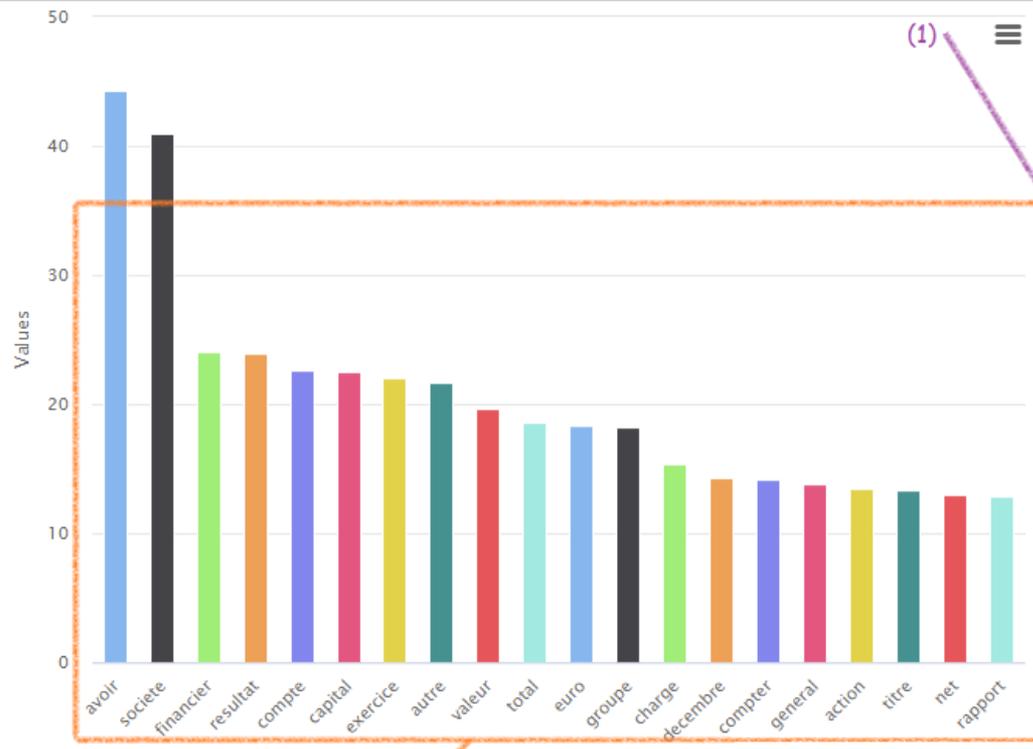
Terms extension  - None  +

Query Reset

Groupings

- Groups
- 1- cible
    - ALL
    - B2B [5804]
    - B2B-B2B2C [10]
    - B2B-B2C [1011]
    - B2C [1295]

TOP KEYWORDS



(1) [Menu icon]

(3) [Legend icon]

(2) [Highcharts.com watermark]

ANALYSES

Documents Tables

- Document properties
- Correlation analyses
- Top Keywords

Parameters

Analysis  Simple  Advanced

Vocabulary

Terms limit

Terms offset

Visualization

Go

Results

Agg. time (s)	7.907
Exp. time (s)	0.023
Result count	4782

- Highlights
- Scoring
- Links Analysis

EXPLORATORY TASKS

Terms filtering

Terms

+ matching

- matching

Parameters

Strictness: Any, All

Fuzzy search: Yes, No

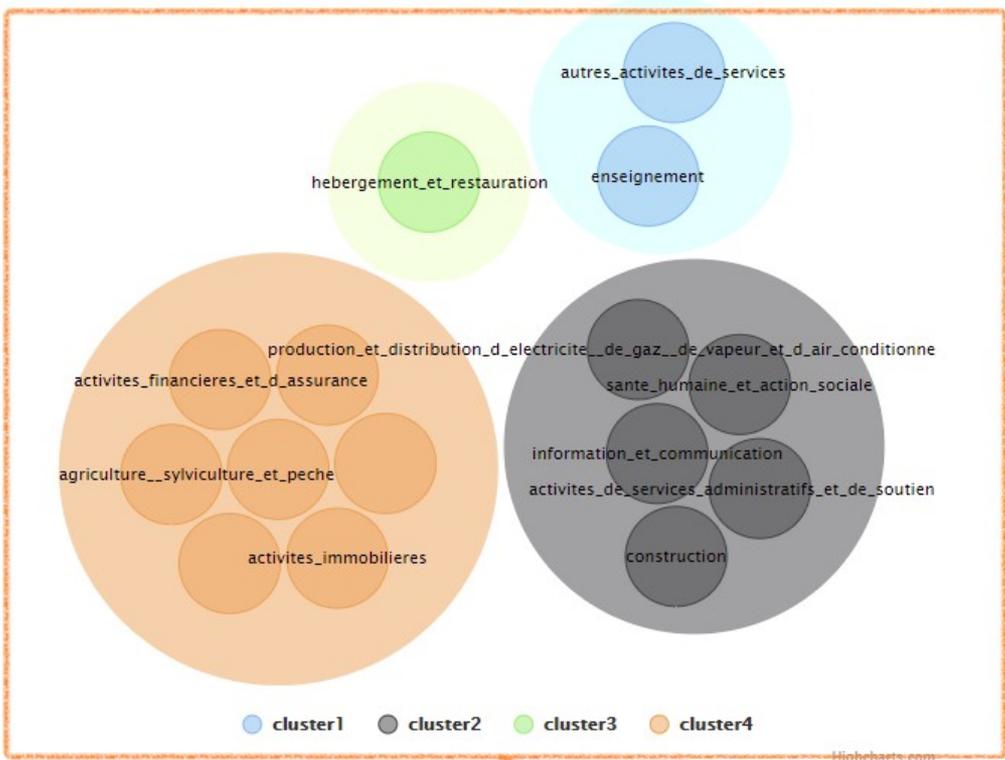
Terms extension: None

Query, Reset

Groupings

- 1- cible
- 2- digitalNativity
- 3- docCategory
- 4- entreprise
- 5- language
- 6- mimeType
- 7- month

TABLE DATA CLUSTERING



(1)

(3)

(2)

ANALYSES

Documents, Tables

Clustering

Parameters

Query type: Simple, Custom

Main table: scores\_entre

Score\_relation, Score\_service, Score\_digital, Entreprise

Join type: INNER JOIN

secteurs\_ac

Join Table: sous\_secteur, secteur\_activite, Entreprise

Group-by: secteurs\_ac

Aggregation: AVG

scores\_entreprises, scores\_entreprises, scores\_entreprises

Analysis: KMeans

KMeans Nb-Classes: 4

# Plan

- ✓ Définitions
- ✓ Entrepôts et lacs de données
- ✓ Métadonnées et gestion des métadonnées
- ✓ Architectures et technologies pour les lacs
- ✓ Discussion, travaux de recherche

