Lacs de données et métadonnées

Séminaire Master 2 IDSM-Kharkiv

Année 2023-2024

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

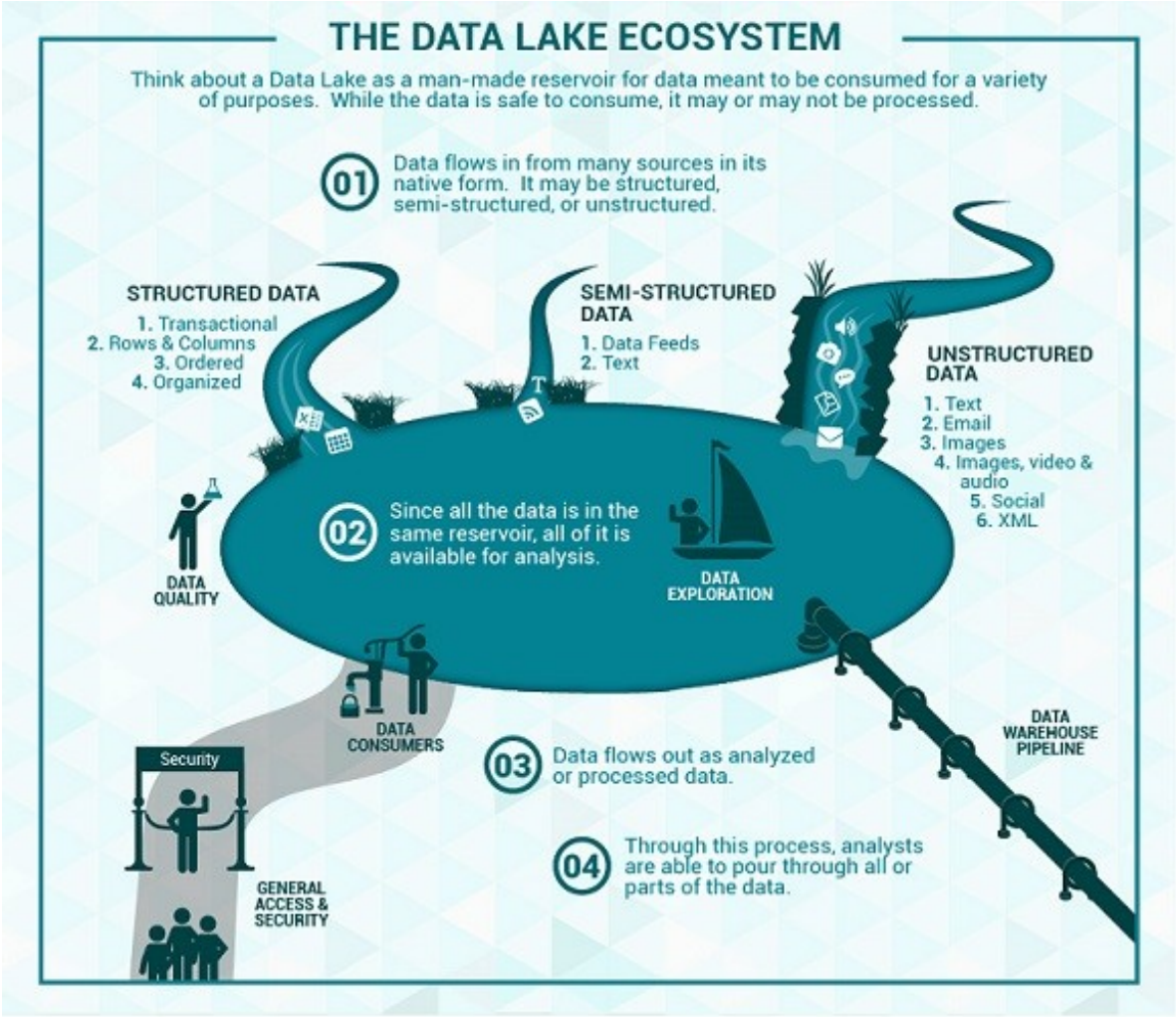
Définition de James Dixon (2010)

“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

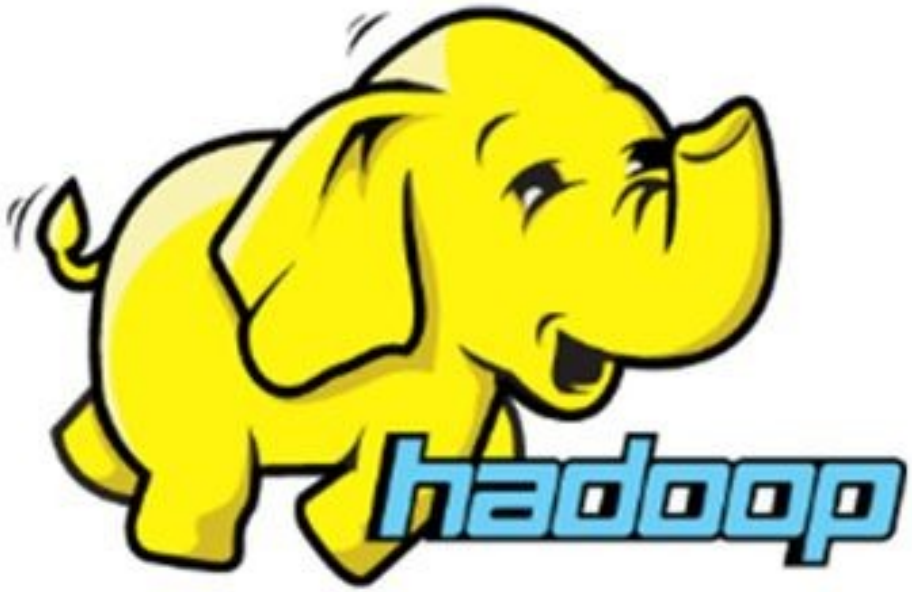
Plus en détail



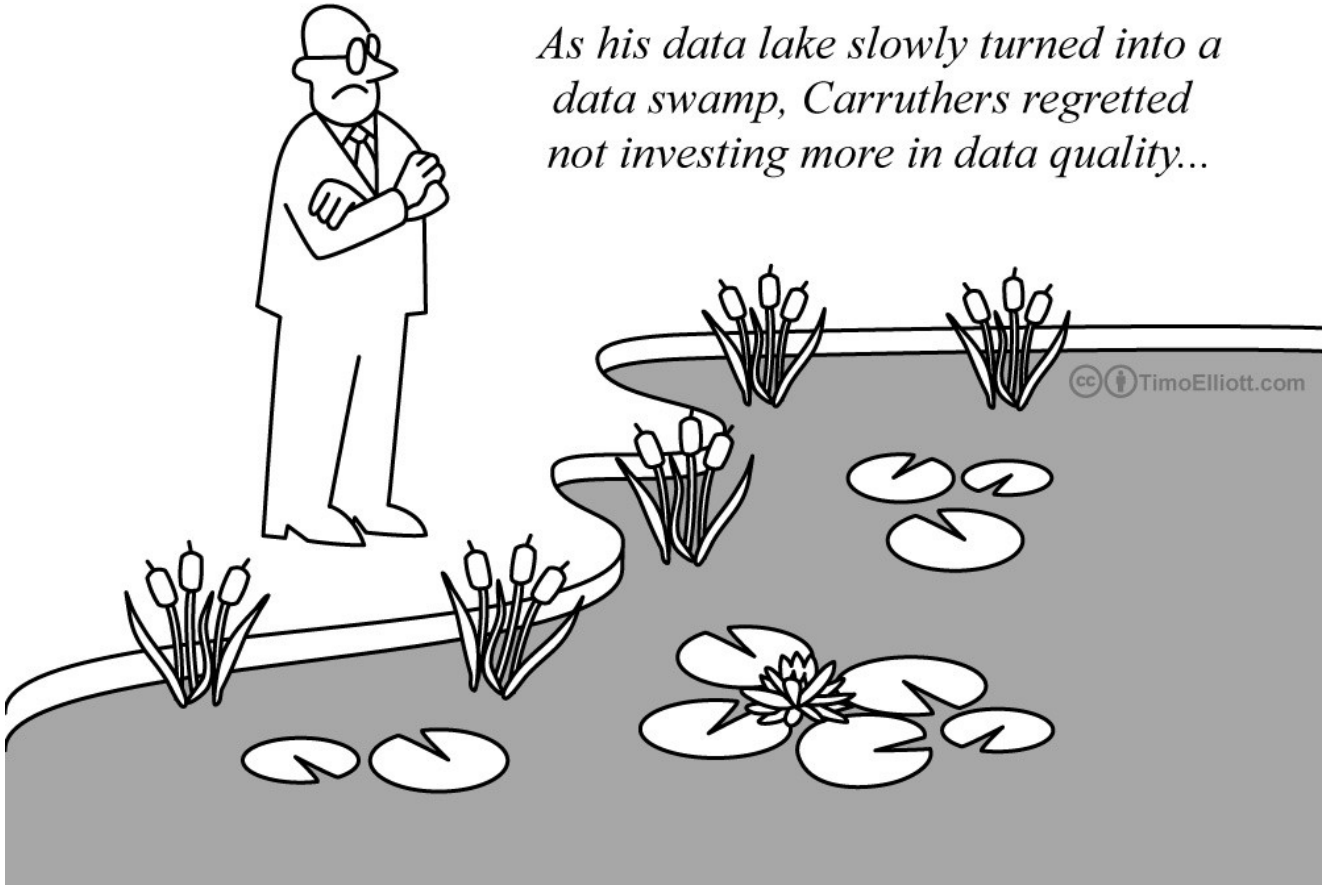
dunnsolutions.com

Définition initiale de l'industrie

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



Danger 1 : le *data swamp*

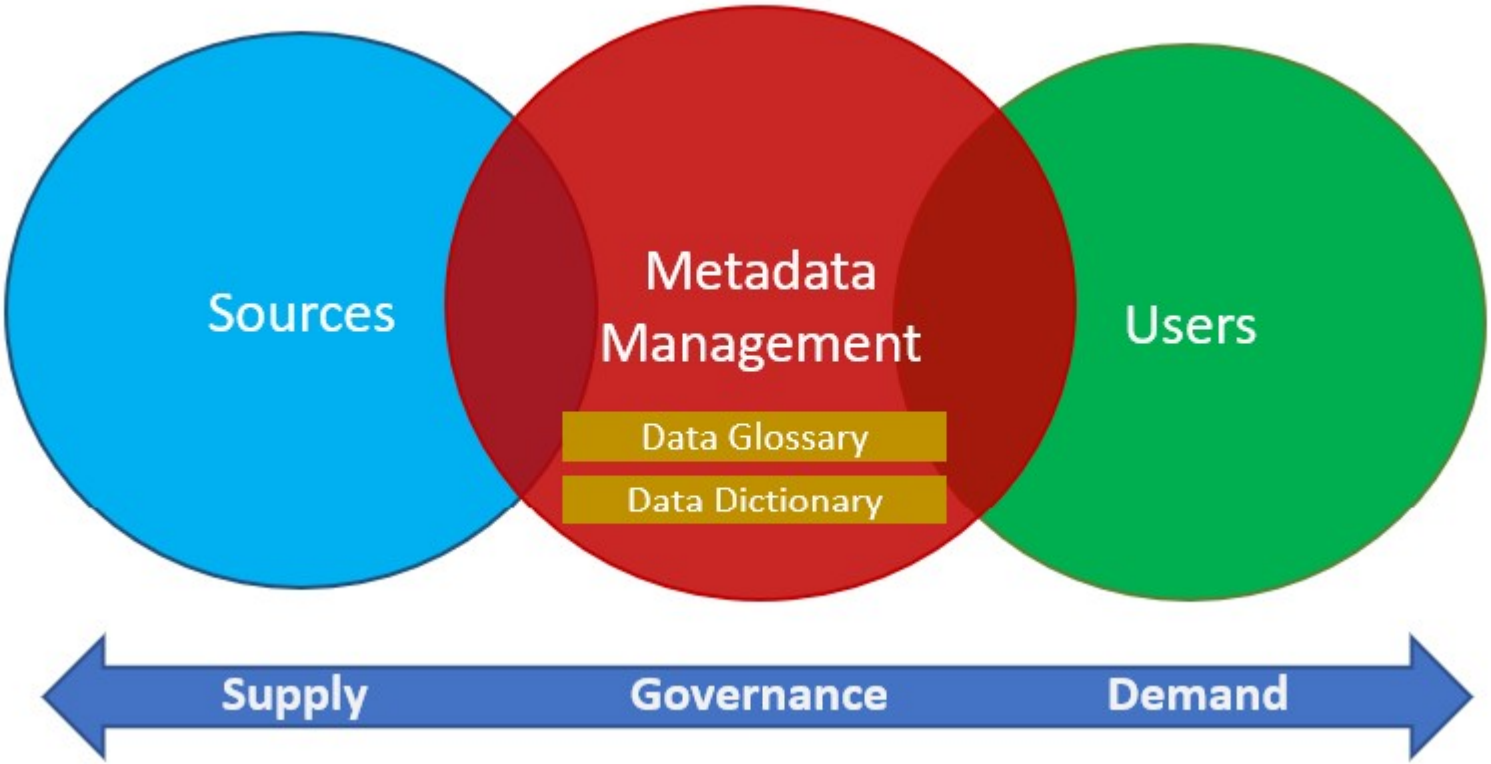


timoelliott.com

Danger 2 : le *data dump*

www.cartoonstock.com

Comment éviter ces dangers ?



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**

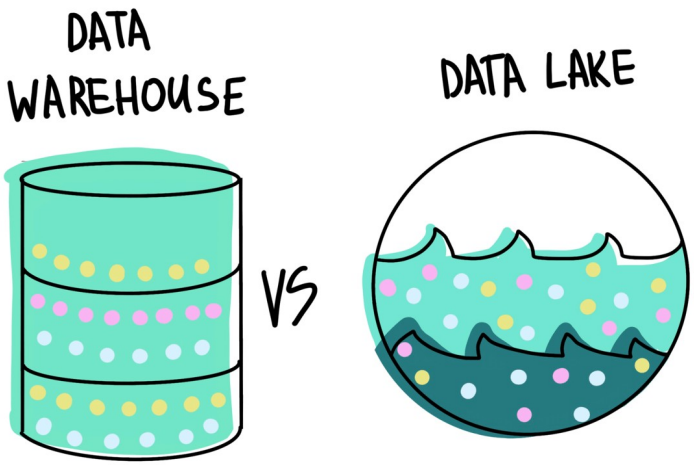
Utilisé **pas seulement** par des *data scientists* et *data analysts*

(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

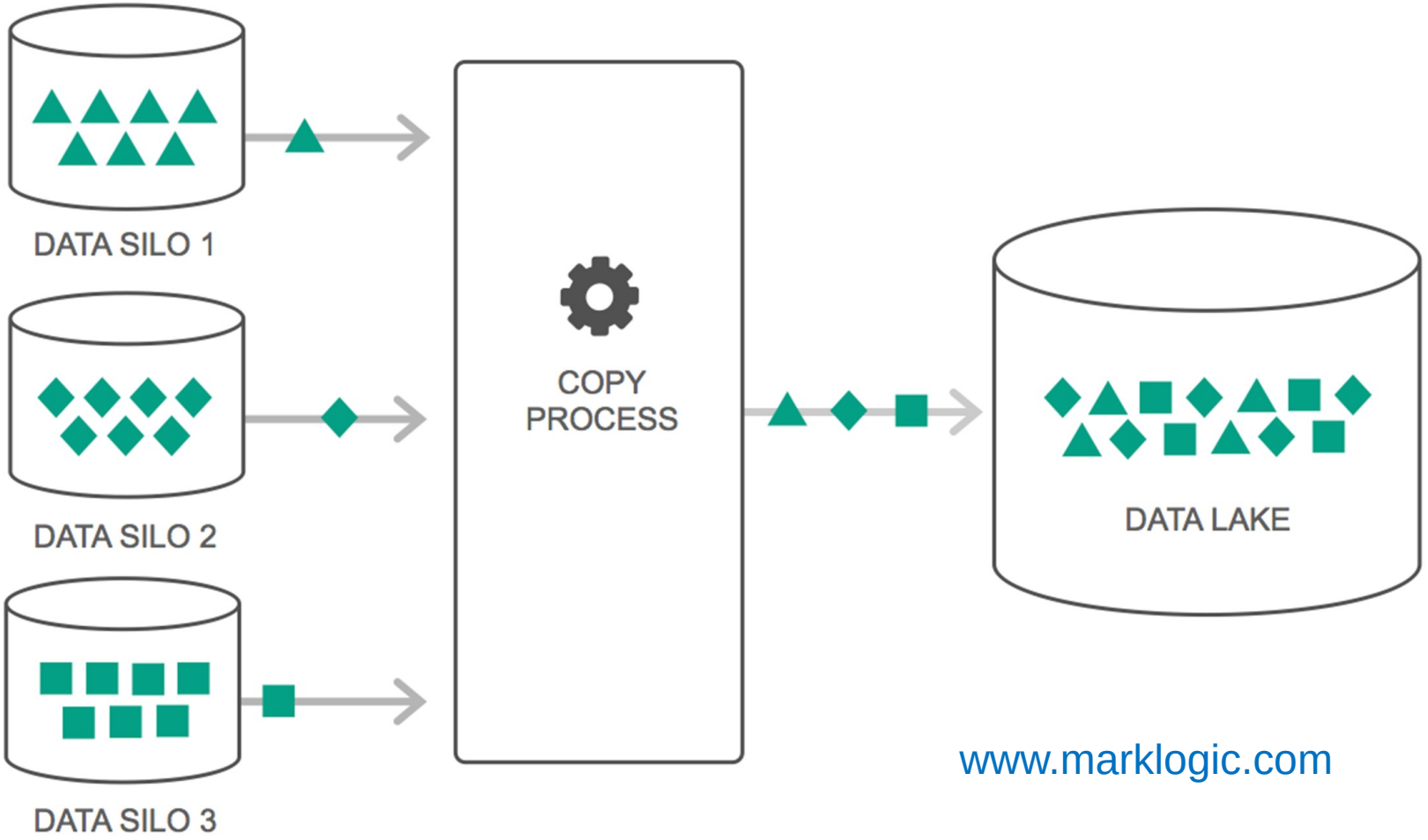


- ✓ Définitions
- Entrepôts et lacs de données
- Métadonnées et modèles de métadonnées
- Architectures et technologies pour les lacs
- Discussion, travaux de recherche

Data lake *vs.* Data warehouse

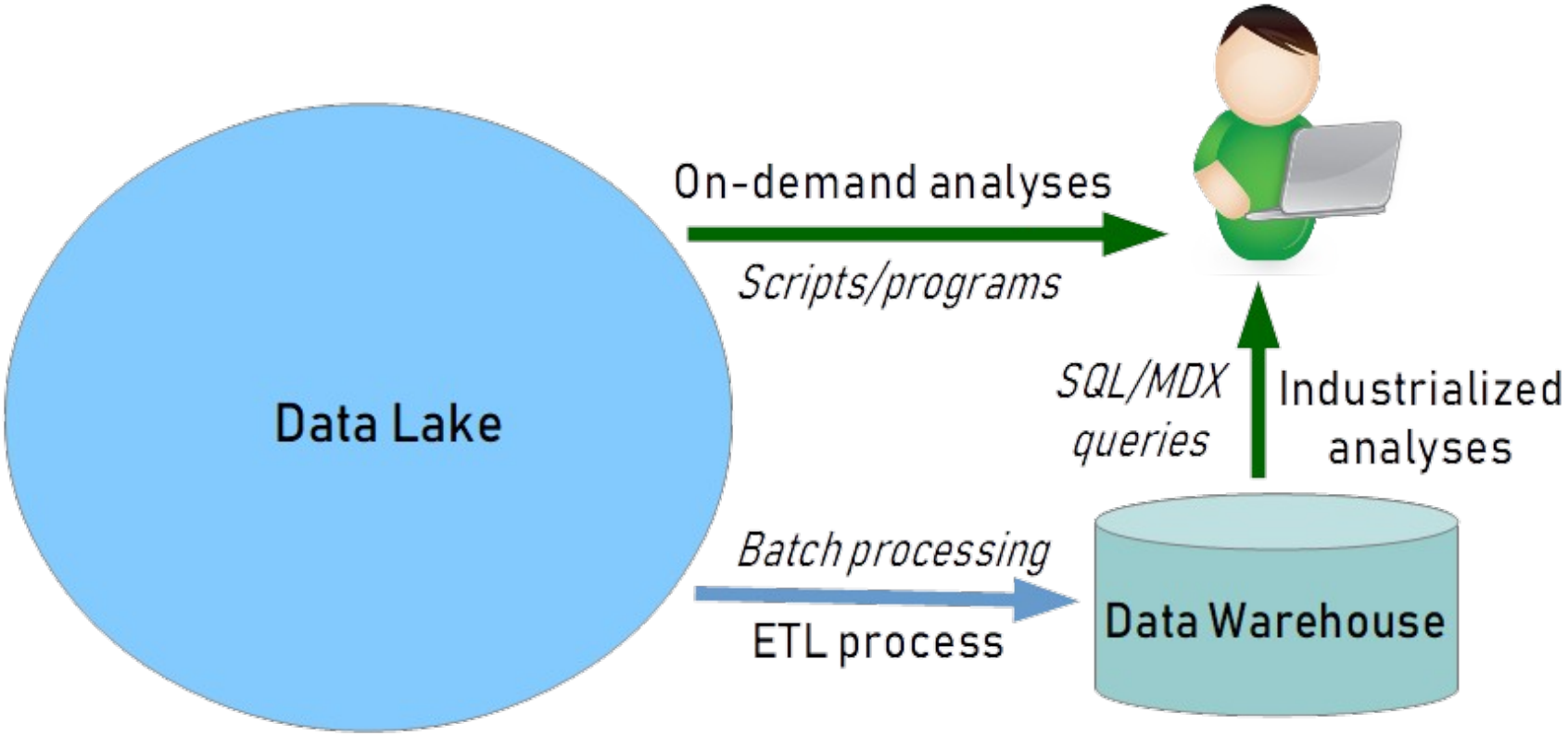
Domaine	Entrepôt de données	Lac de données
Données	Nettoyées	Brutes
Schéma	*Schema on write*	*Schema on read*
Sources de données	En nombre limité	Multiples
Structure	Données structurées	Données structurées, semi-structurées, non-structurées
Analyses	Industrialisées	À la demande
Accès aux données	SQL/MDX	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	*Data scientists*

Entrepôt comme source d'un lac



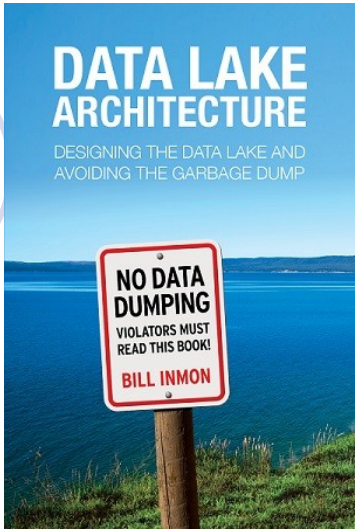
www.marklogic.com

Lac comme source d'un entrepôt



Sawadogo et Darmont 2019

Entrepôt dans un lac

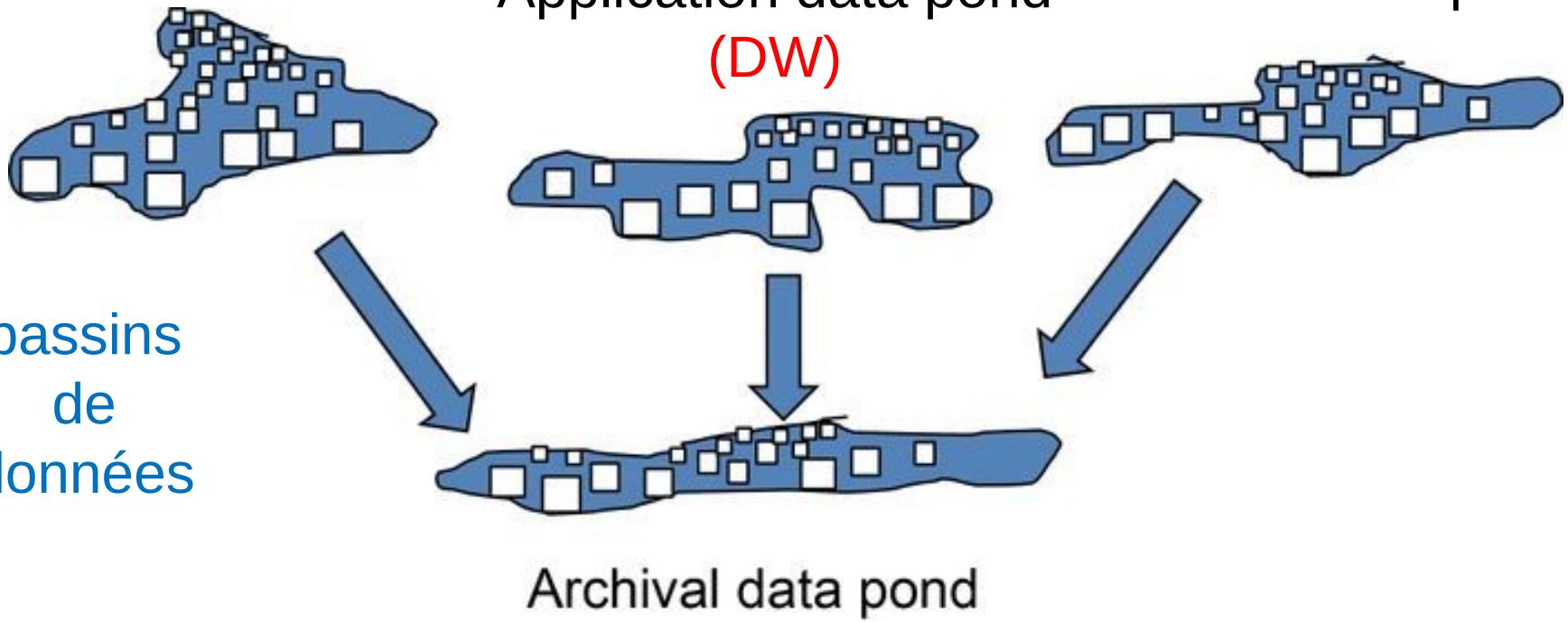


(Raw data pond)

Analog data pond

Application data pond
(DW)

Textual data pond



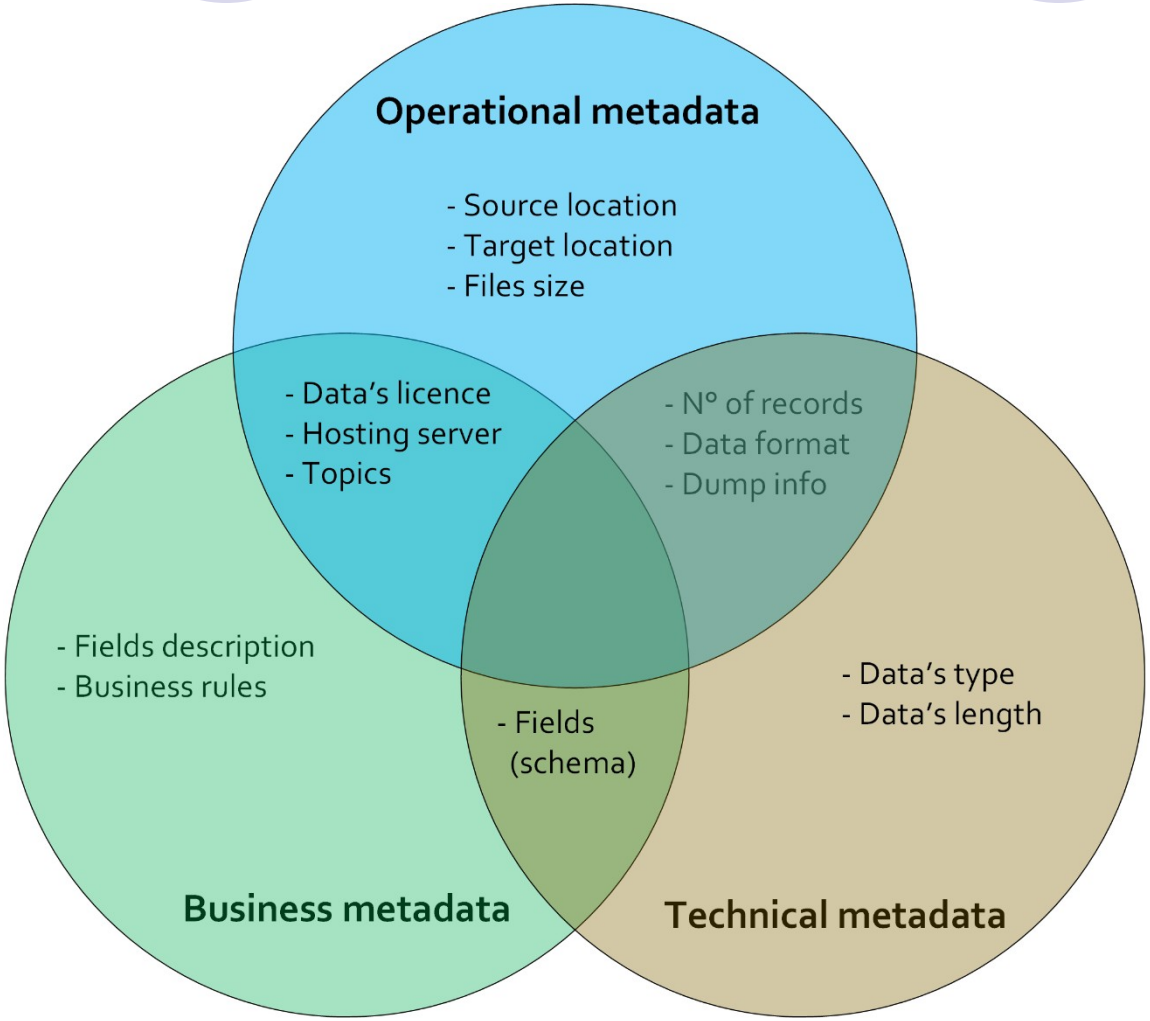
bassins
de
données

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…)

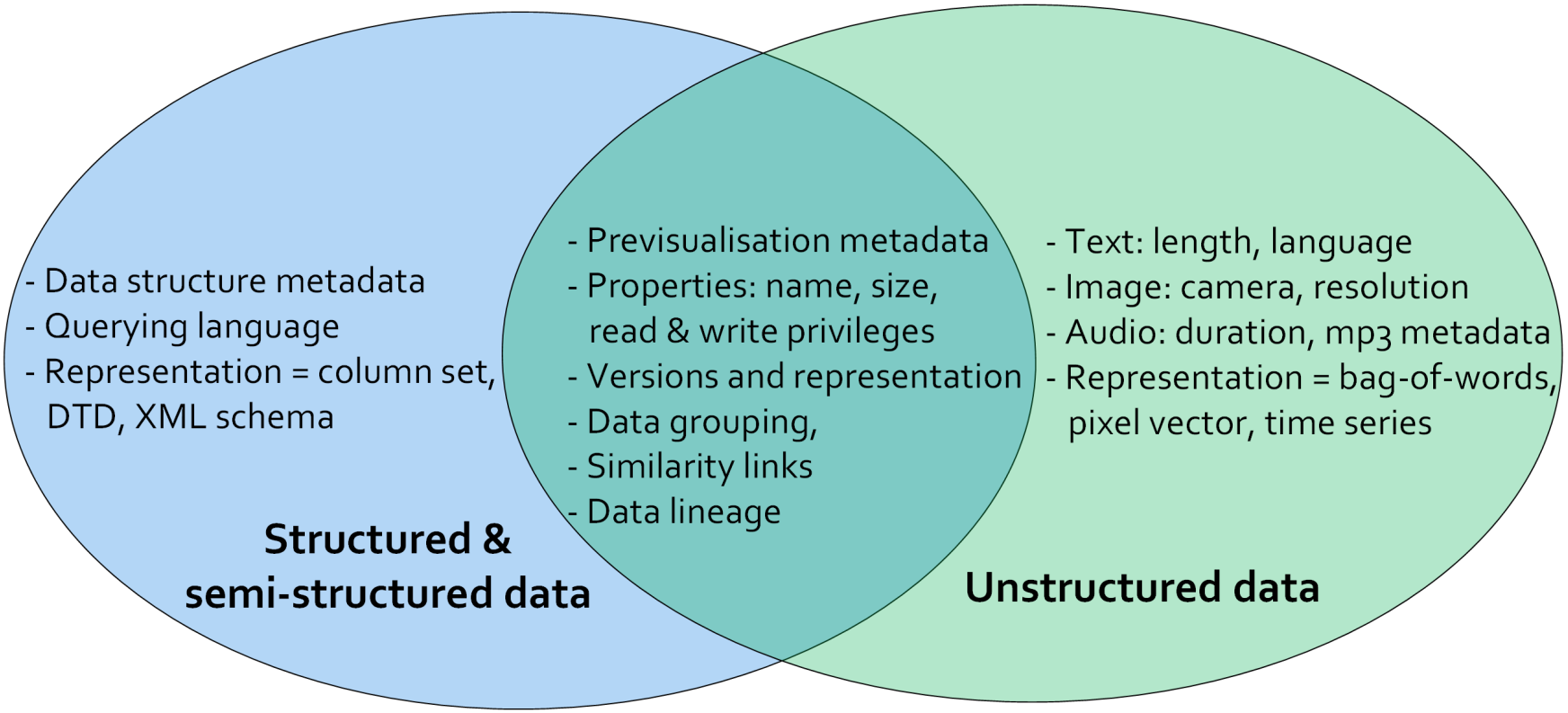
Métadonnées intra-objet	Exemples
Propriétés	Nom de fichier, taille, date de création…
Prévisualisations/résumés	Schéma, nuage de mots...
Versions et représentations	Transformation des données
Métadonnées sémantiques	Description, catégorie...

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	*Logs*

Typologie opérationnelle

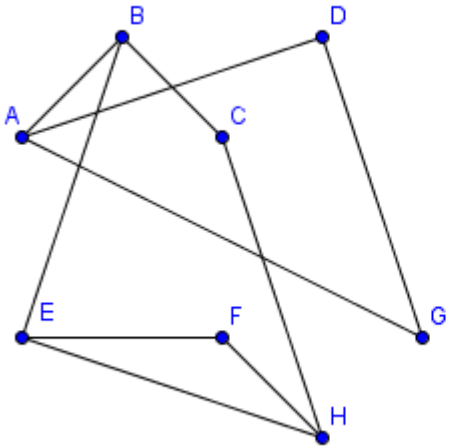


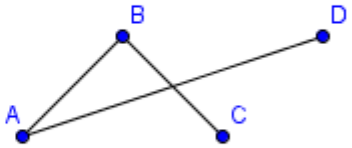
Sawadogo et Darmont 2019

Représentation logique des métadonnées

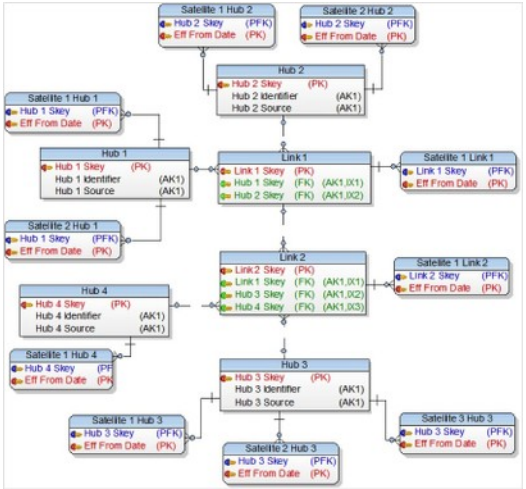
Modèles relationnels ou NoSQL

Graphes

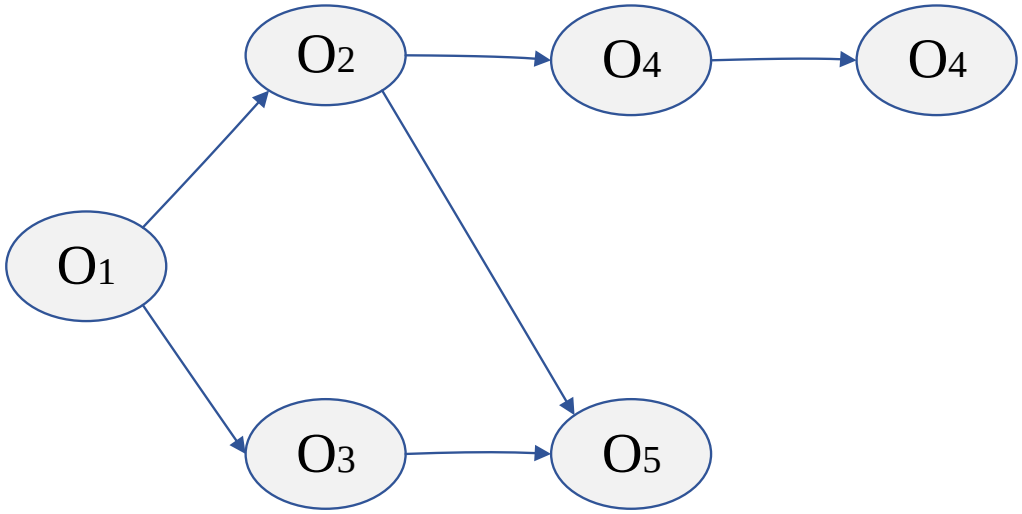




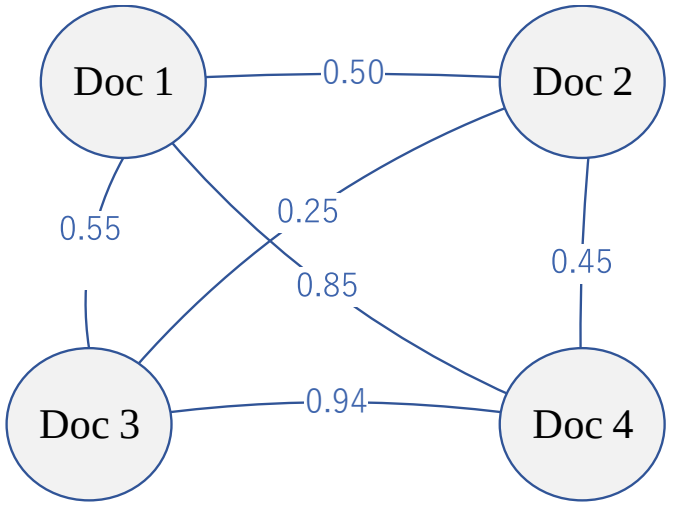
Modélisation ensembliste



Ex. de métadonnées en graphes



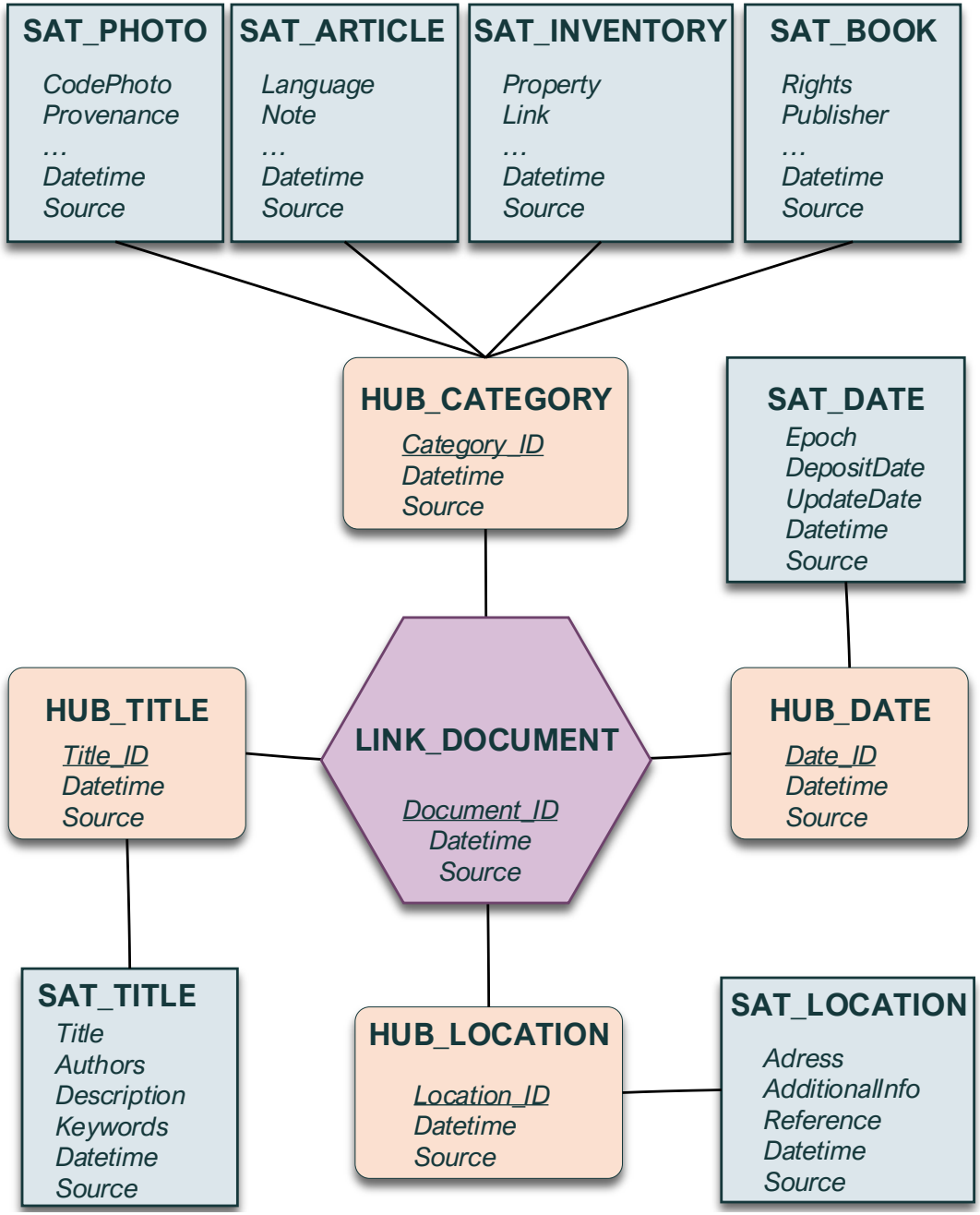
a) Provenance graph



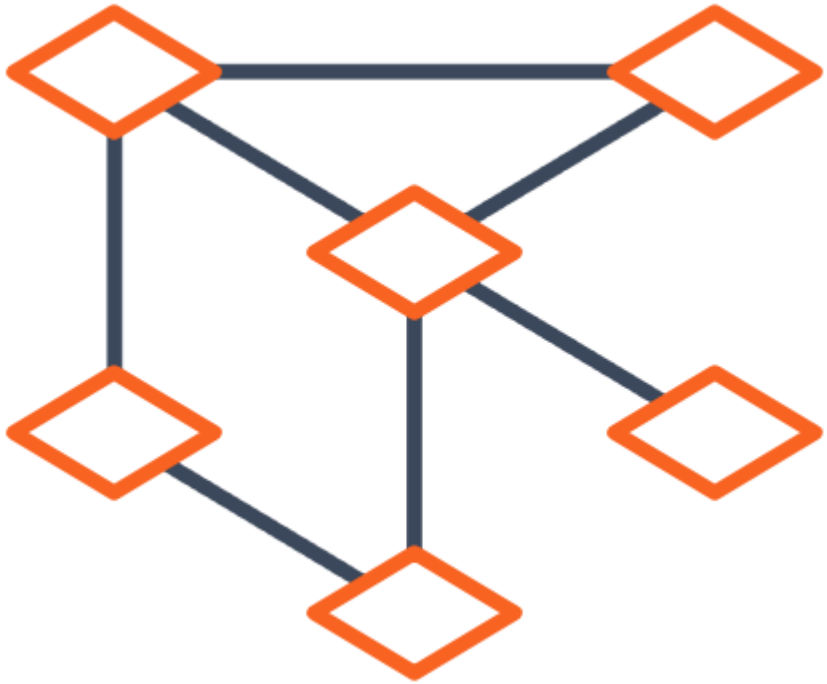
b) Similarity graph

Ex. de métadonnées en data vault

Nogueira et al. 2018

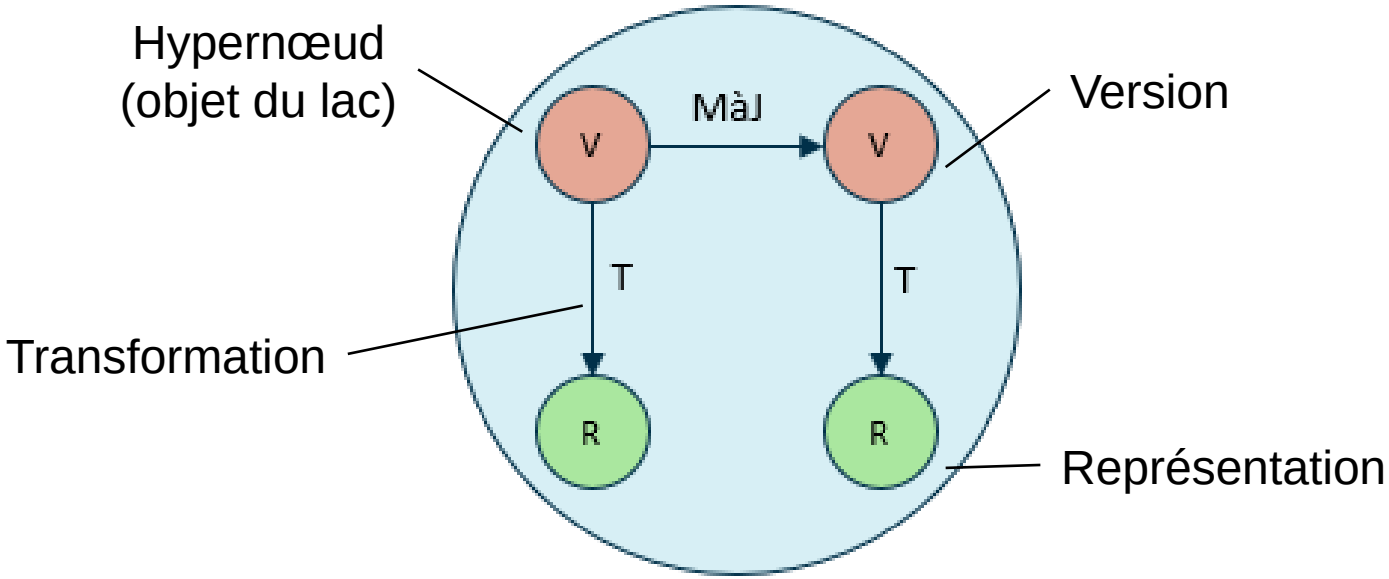


Modèles de métadonnées récents



blog.tensorflow.org

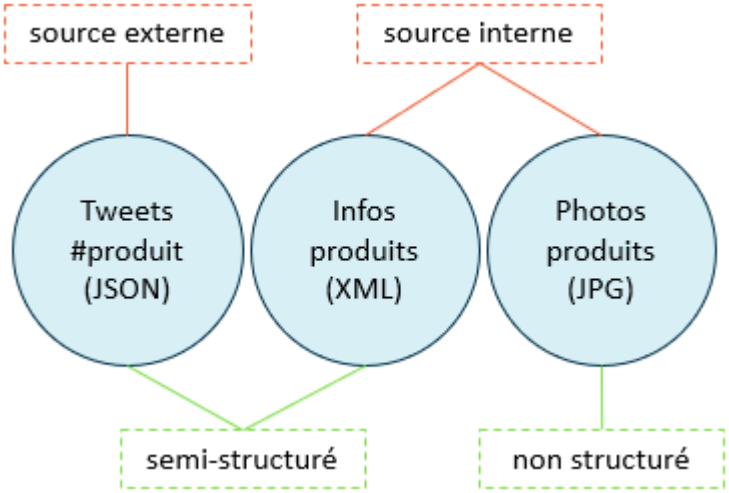
Graphe d'hypernœuds



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

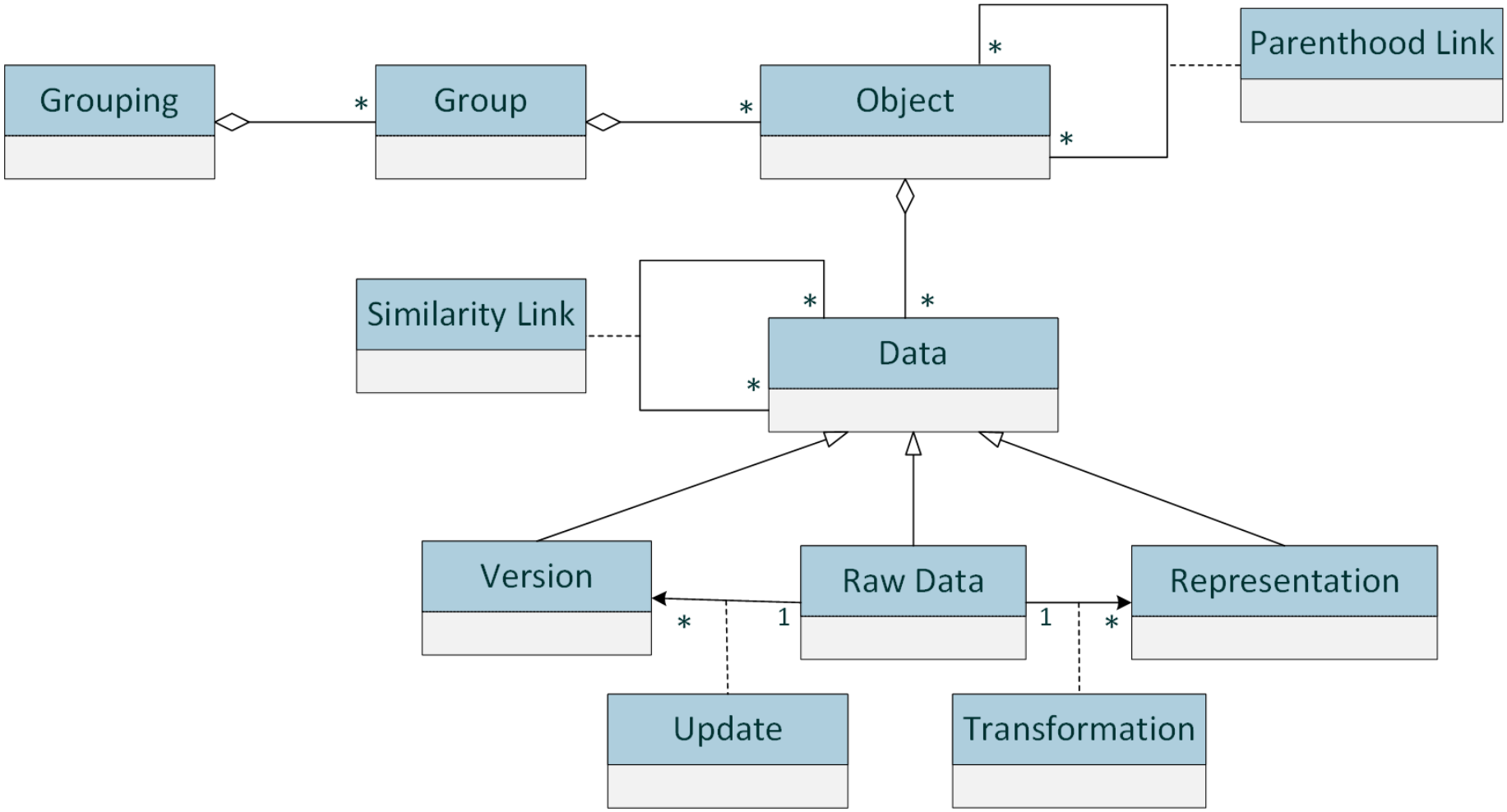
MEDAL

- Métadonnées inter-objets
  - Exemple de regroupements



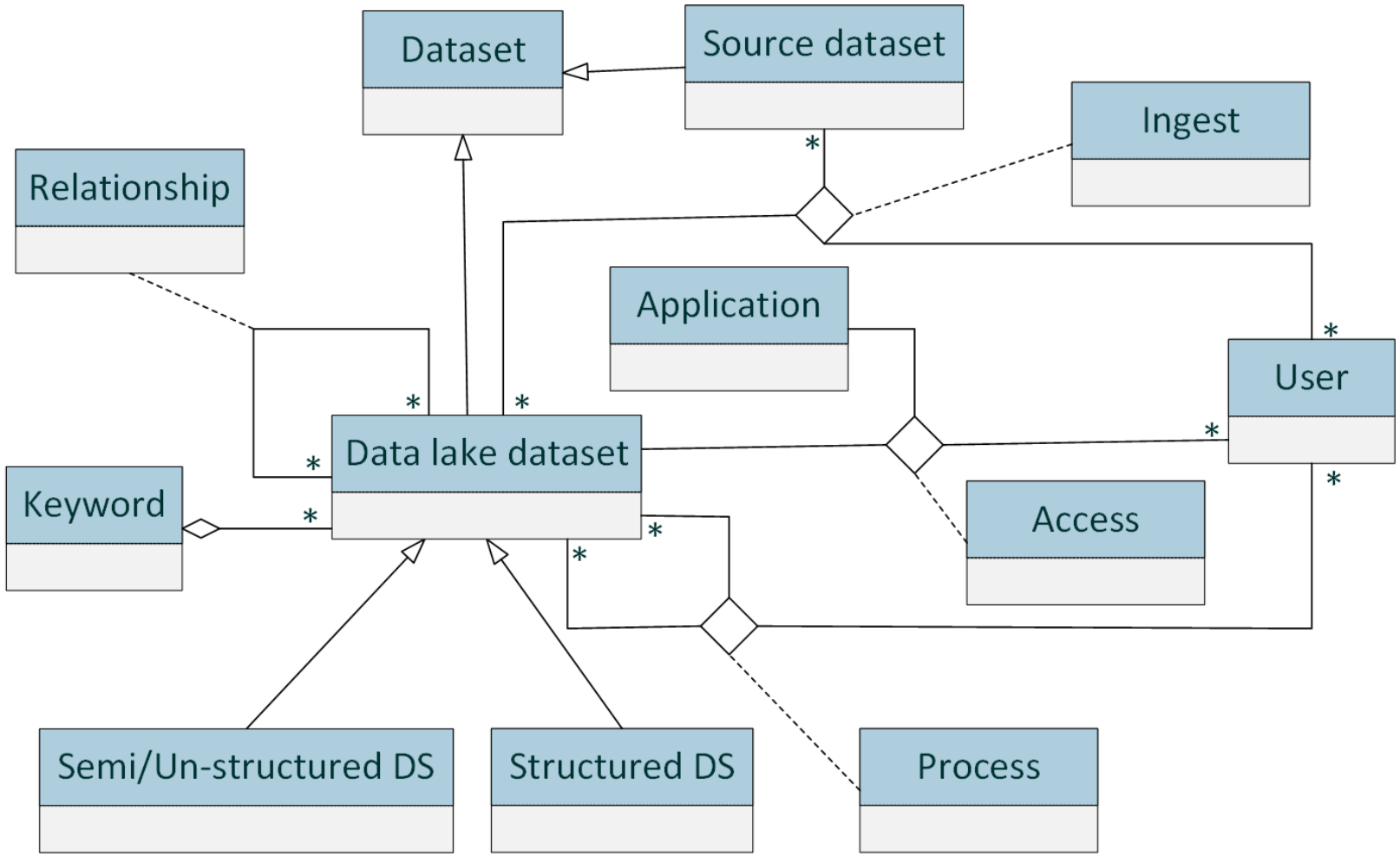
- Métadonnées globales
  - Dans des nœuds non connectés au reste du graphe

MEDAL : Modèle conceptuel

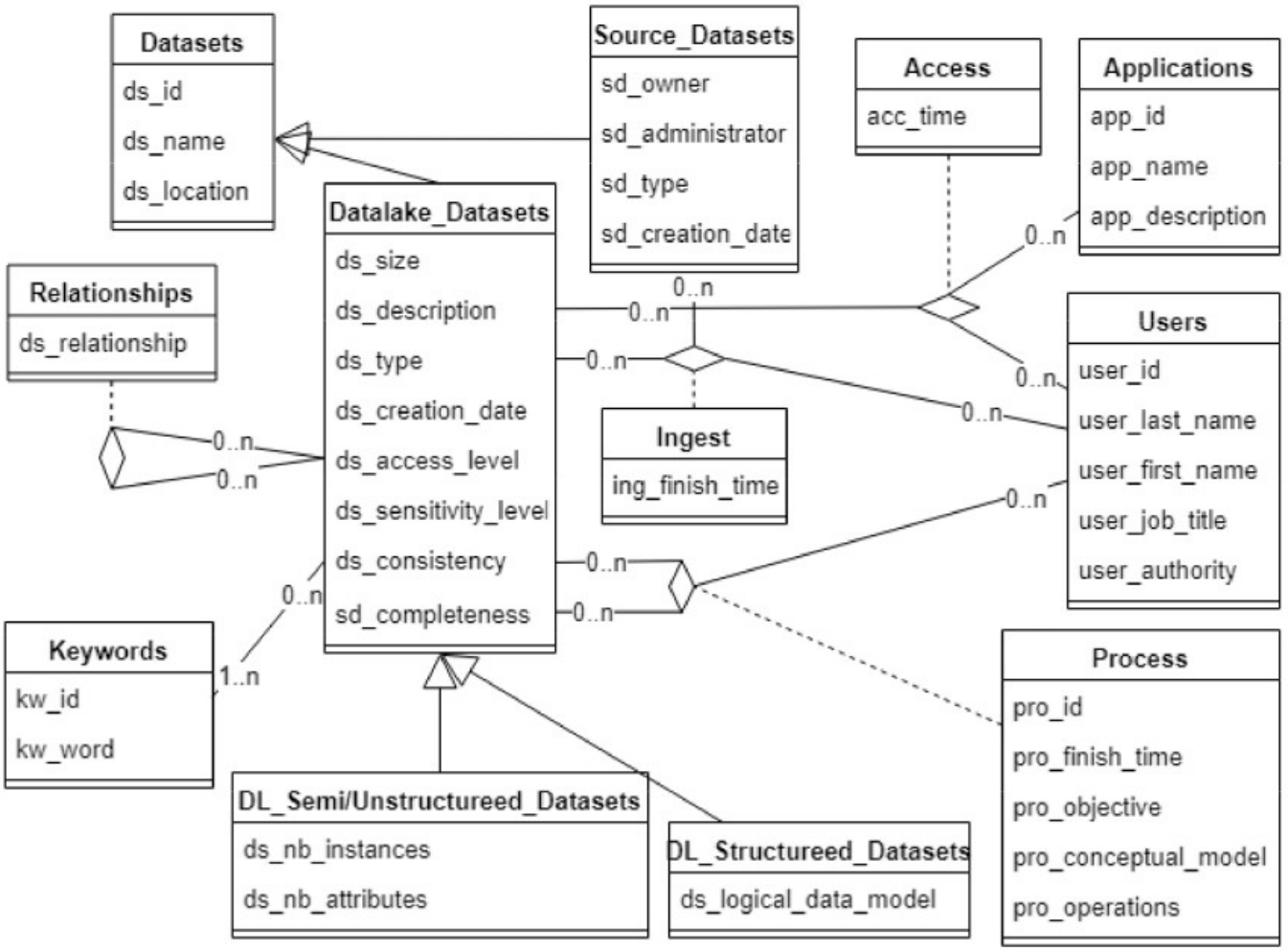


DAMMS

Ravat & Zhao 2019

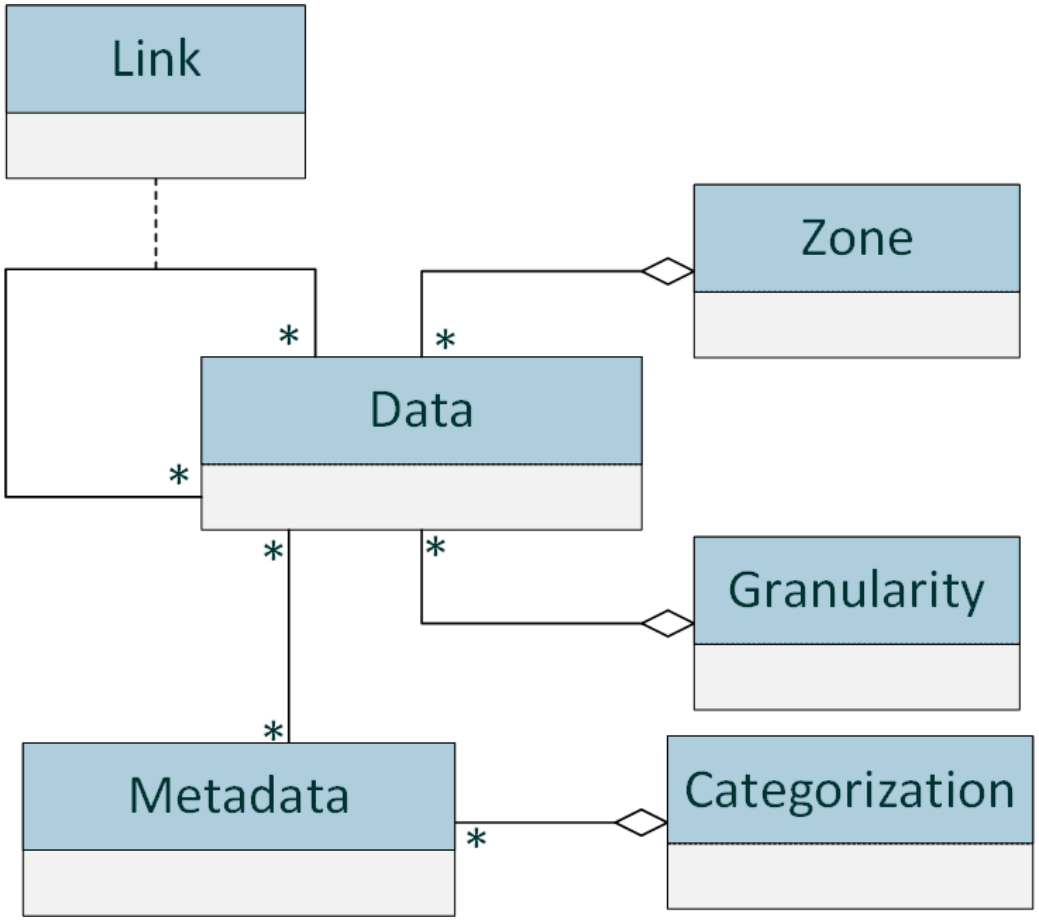


DAMMS dans le détail



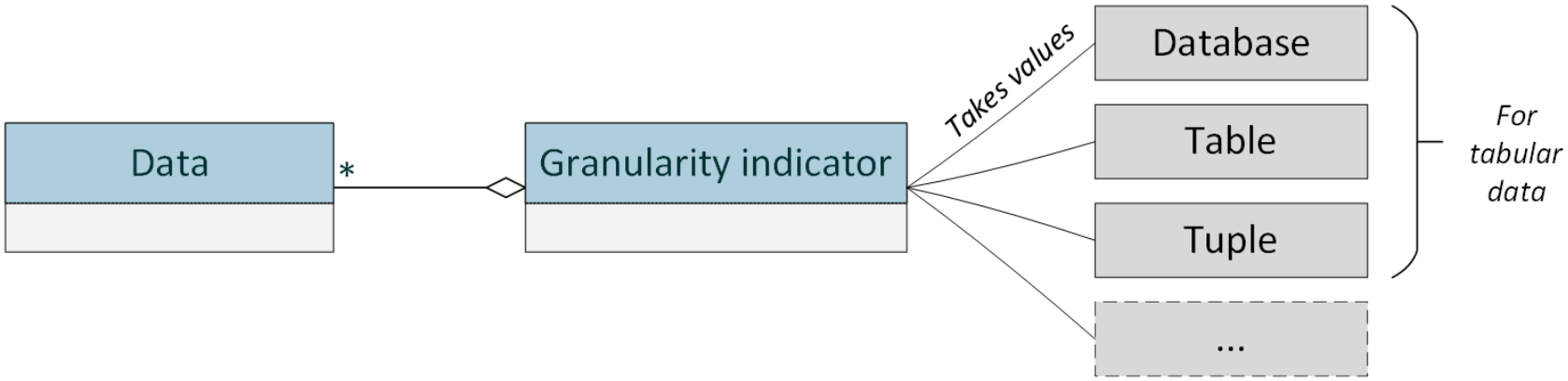
HANDLE

Eichler et al. 2020



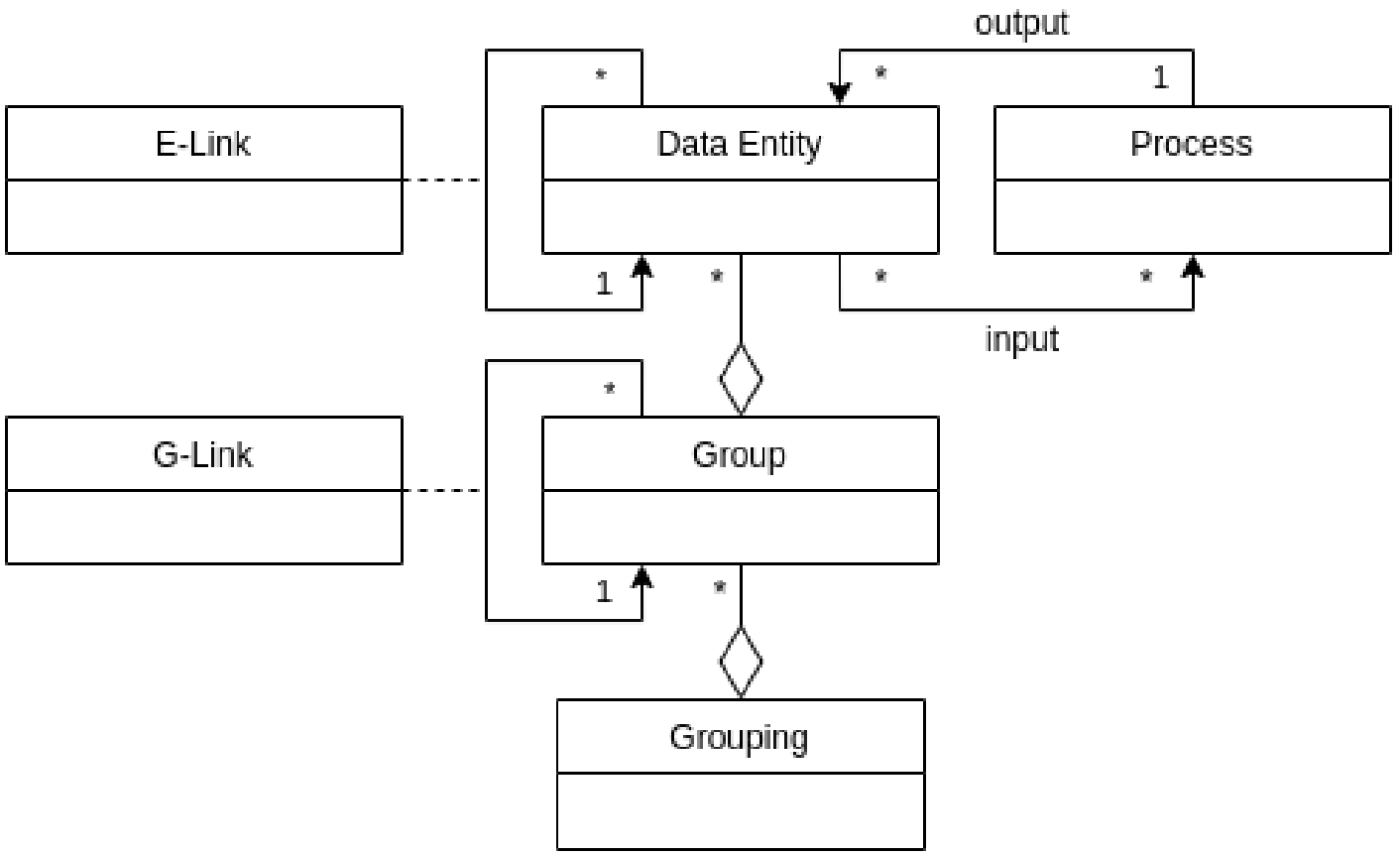
Extensions HANDLE

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



goldMEDAL

Scholly et al. 2021



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

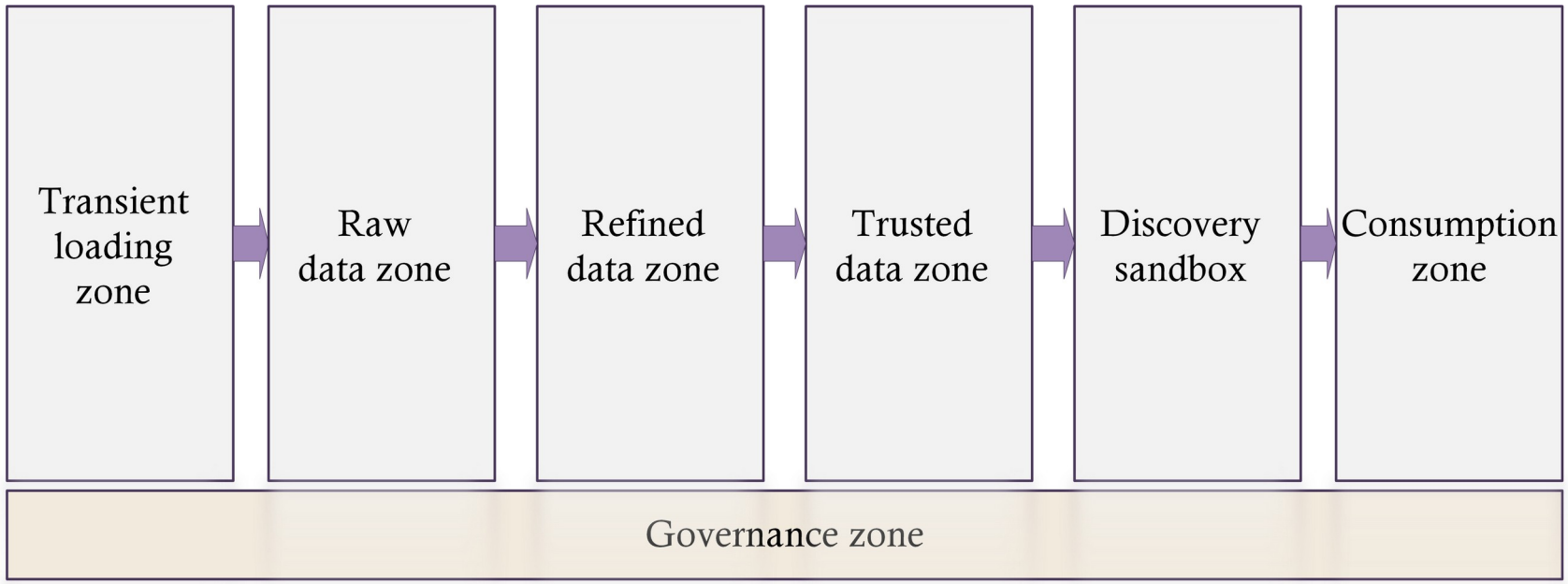
Caractéristiques	MEDAL	DAMMS	HANDLE	goldMEDAL
Enrichissement sémantique	X	X	X	X
Polymorphism/zones multiples	X	X	X	X
Versionnement des données	X	X		X
Suivi des usages	X	X	X	X
Catégorisation	X	X	X	X
Liens de similarité	X	X	X	X
Propriétés des métadonnées	X	X	X	X
Granularités multiples			X	X
Total	7/8	7/8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

Plan

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

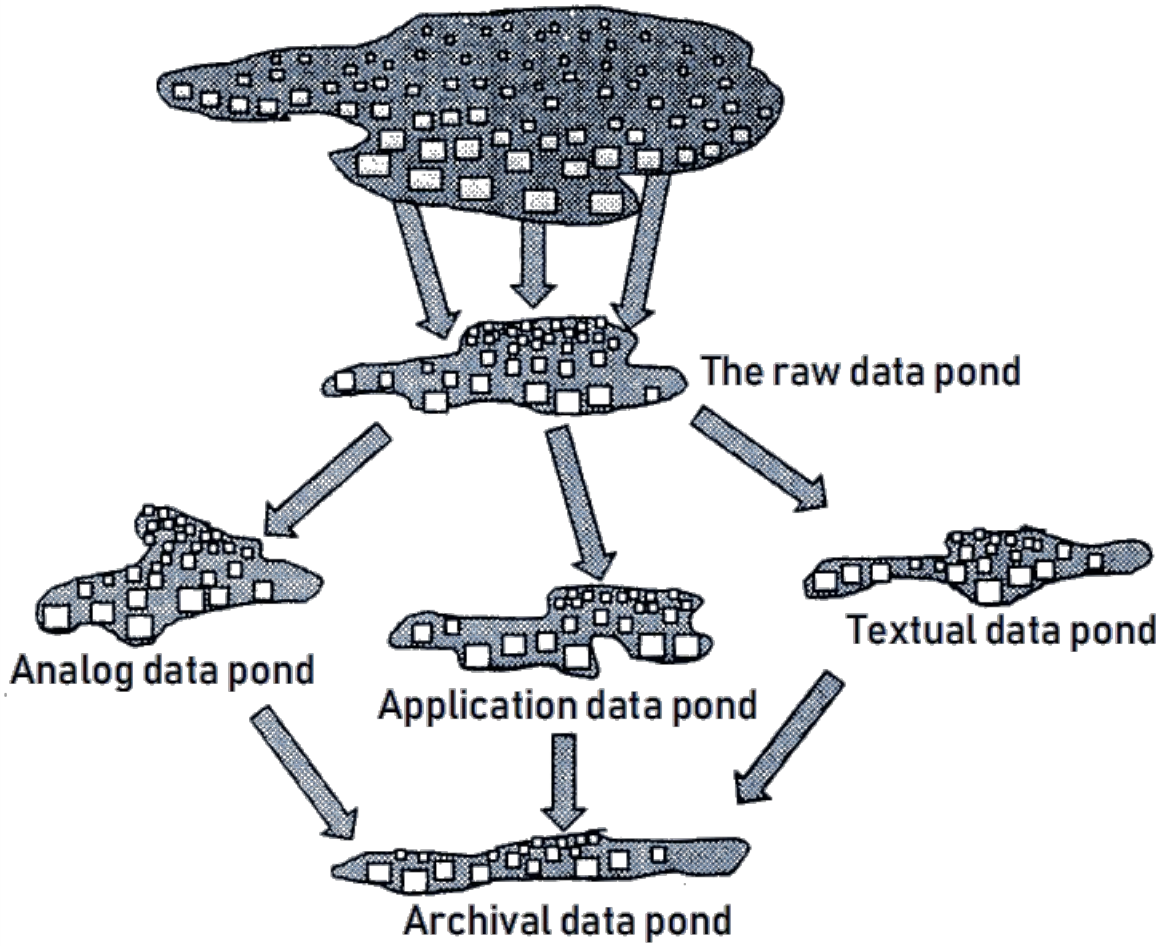
Architecture en zones

Zaloni 2015



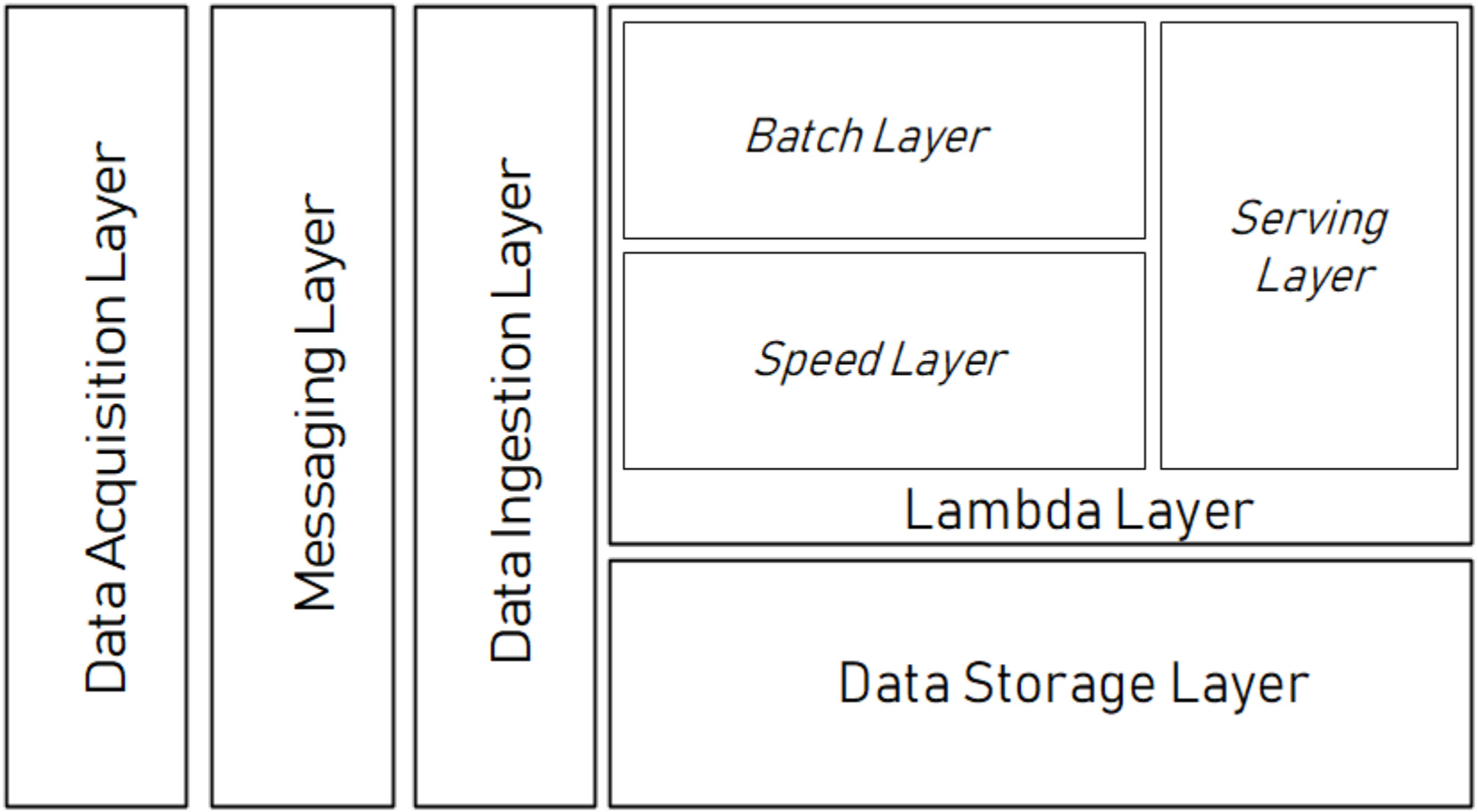
Architecture en bassins

Inmon 2016



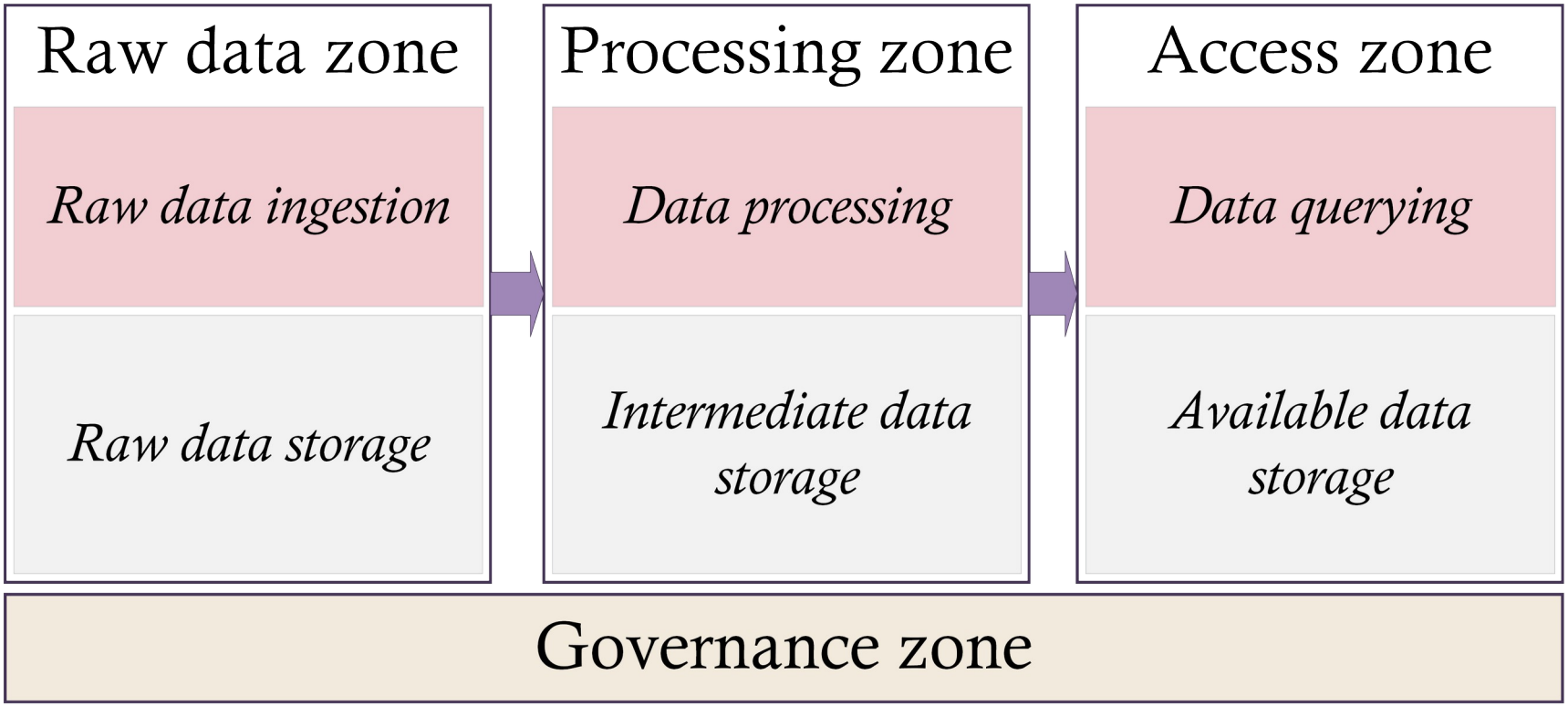
Architecture lambda

John & Misra 2017

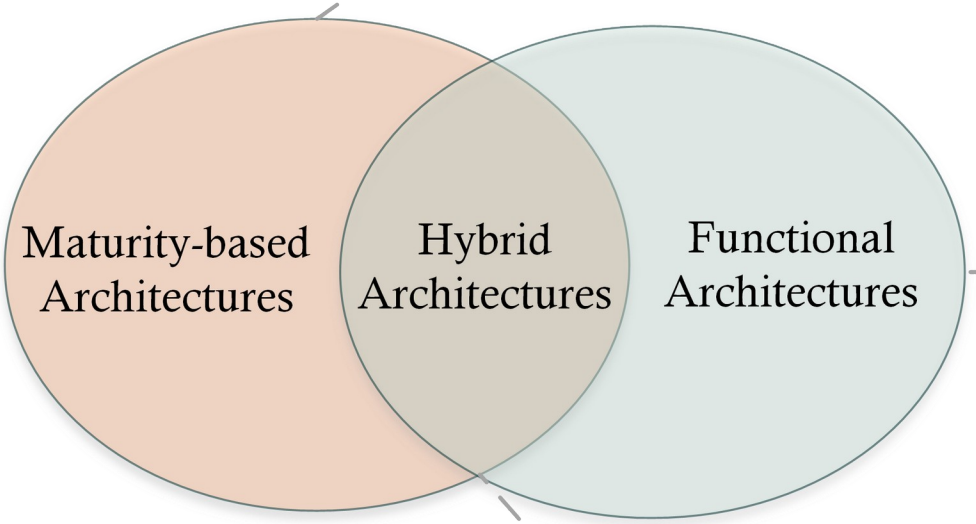


Architecture hybride

Ravat & Zhao 2019



Architectures fonctionnalité × maturité

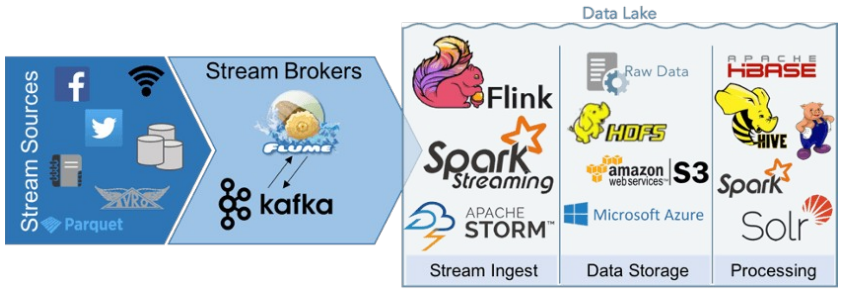


Zikopoulos et al. (2015)
Laplante and Sharma (2016)
Tharrington (2017)
…

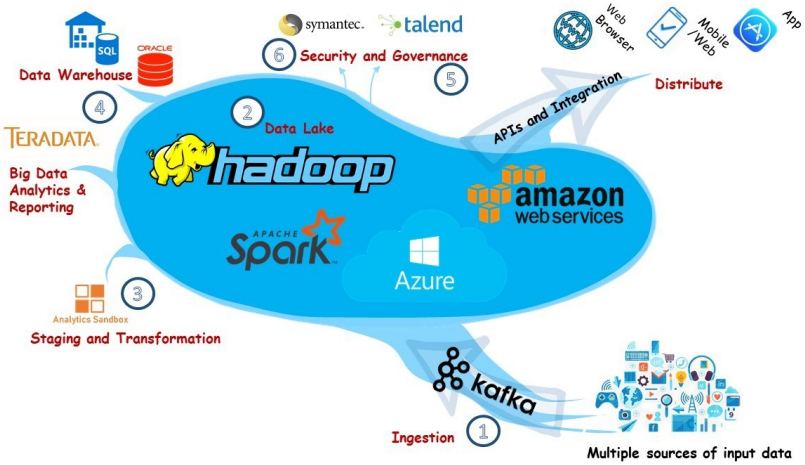
John and Misra (2017)
Quix and Hai (2018)
Mehmood et al. (2019)
…

Inmon (2016)
Ravat and Zhao (2019)
…

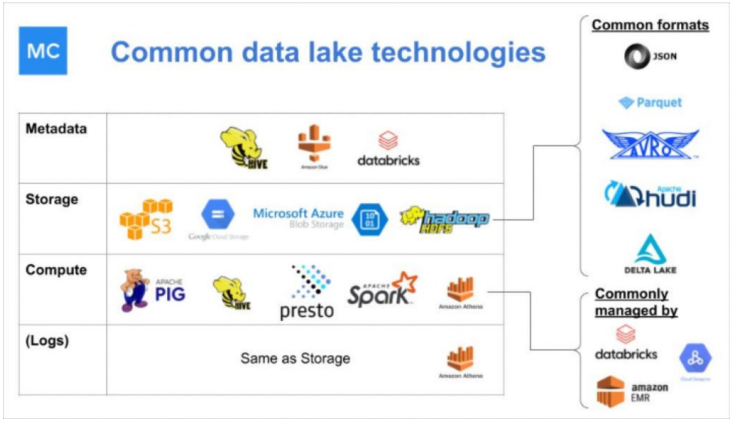
Technos pour les lacs de données



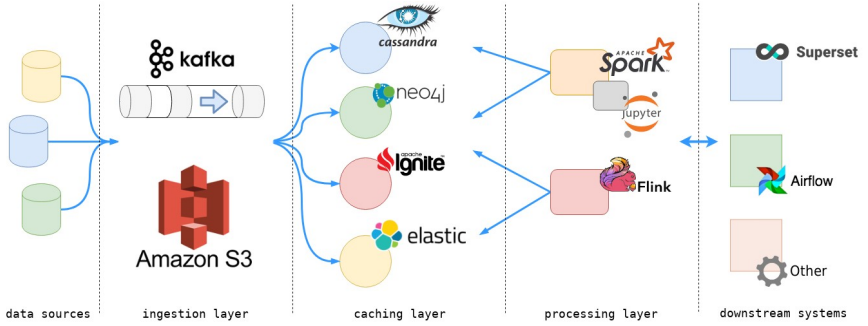
7wdata.be



kms-world.com

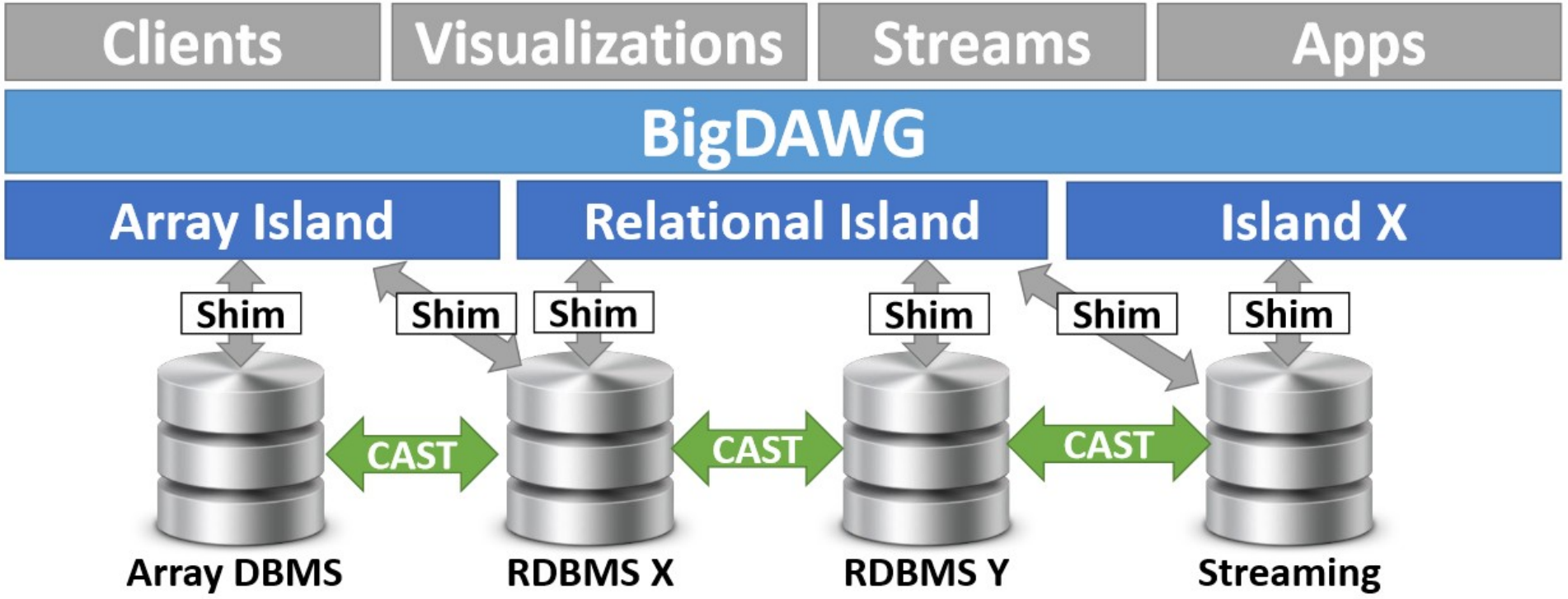


www.montecarlo.com



smartcat.io

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

Les spécialistes

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	Incompatibilité avec des méthodes
Faible coût de stockage	Incohérences de données
Fidélité aux données	Confusion autour du concept de lac
Gestion de données non structurées	Besoin d’interfaces d’accès données
Intégration des données en temps réel	Besoin de métadonnées adaptées
Détection de relations entre données	Manque de standards de gouvernance
Analyses à la volée	Compétences pour la mise en œuvre
Passage à l’échelle	Sécurité en maturation
Tolérance aux pannes	Monolithiques, « challengés » par les
Exploitation de données tierces	maillages de données (*data meshes*)

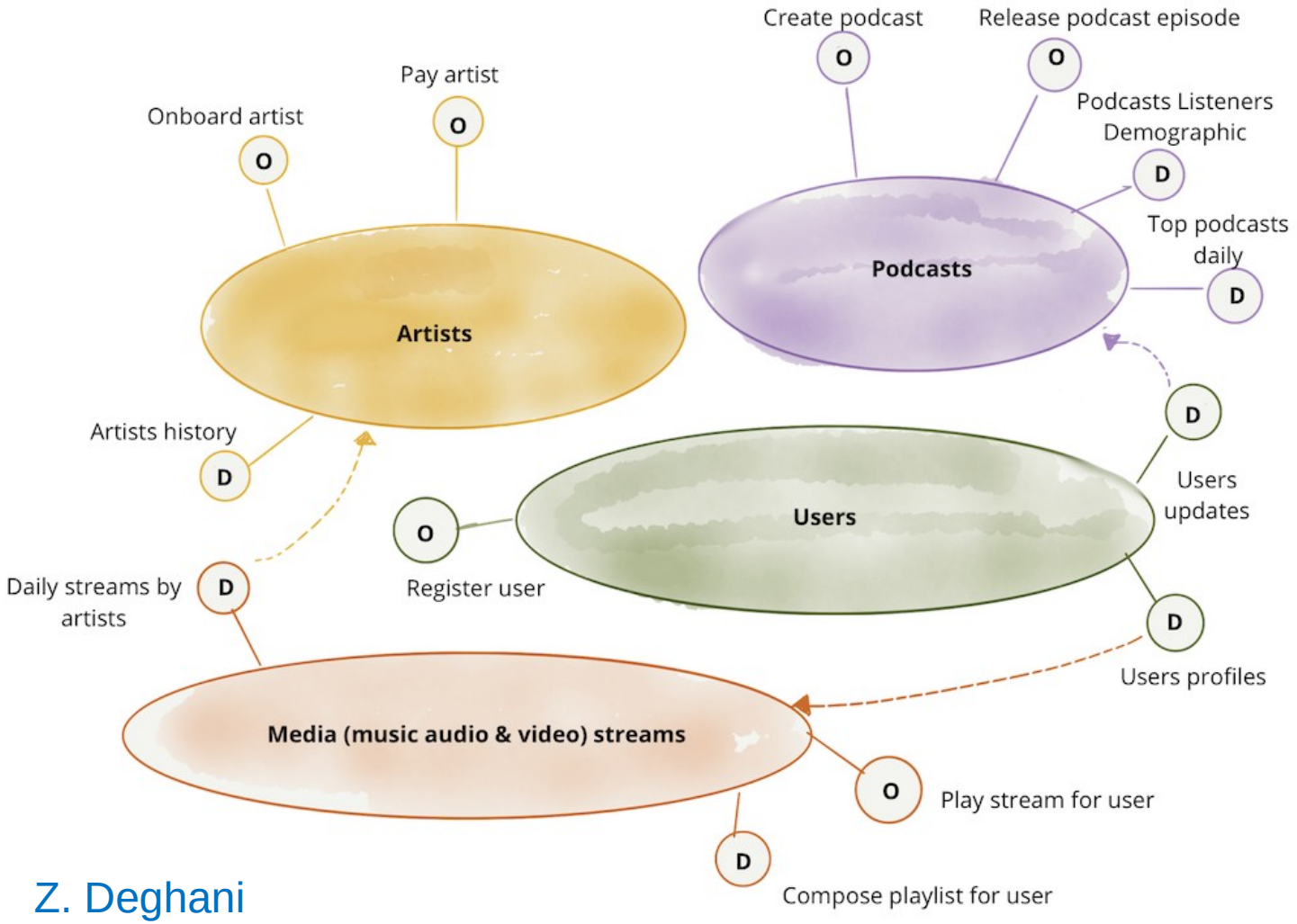
Data mesh ou la décentralisation

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, Accessibles, Interopérables, Réutilisables (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
R. El-Idrissi
2023-2026

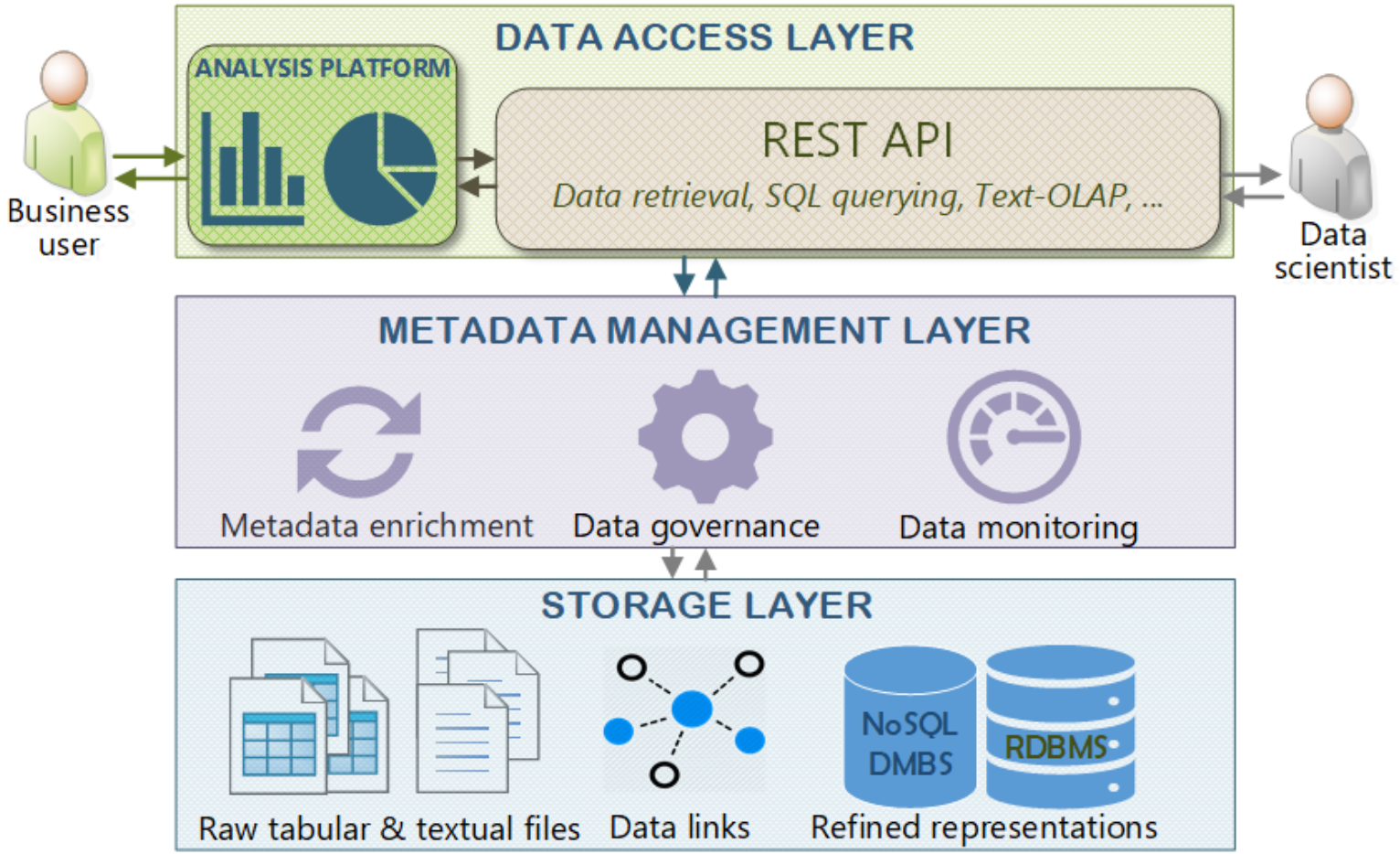
Stage
A. Derder
2023

Stage
R. Aoudj
2023

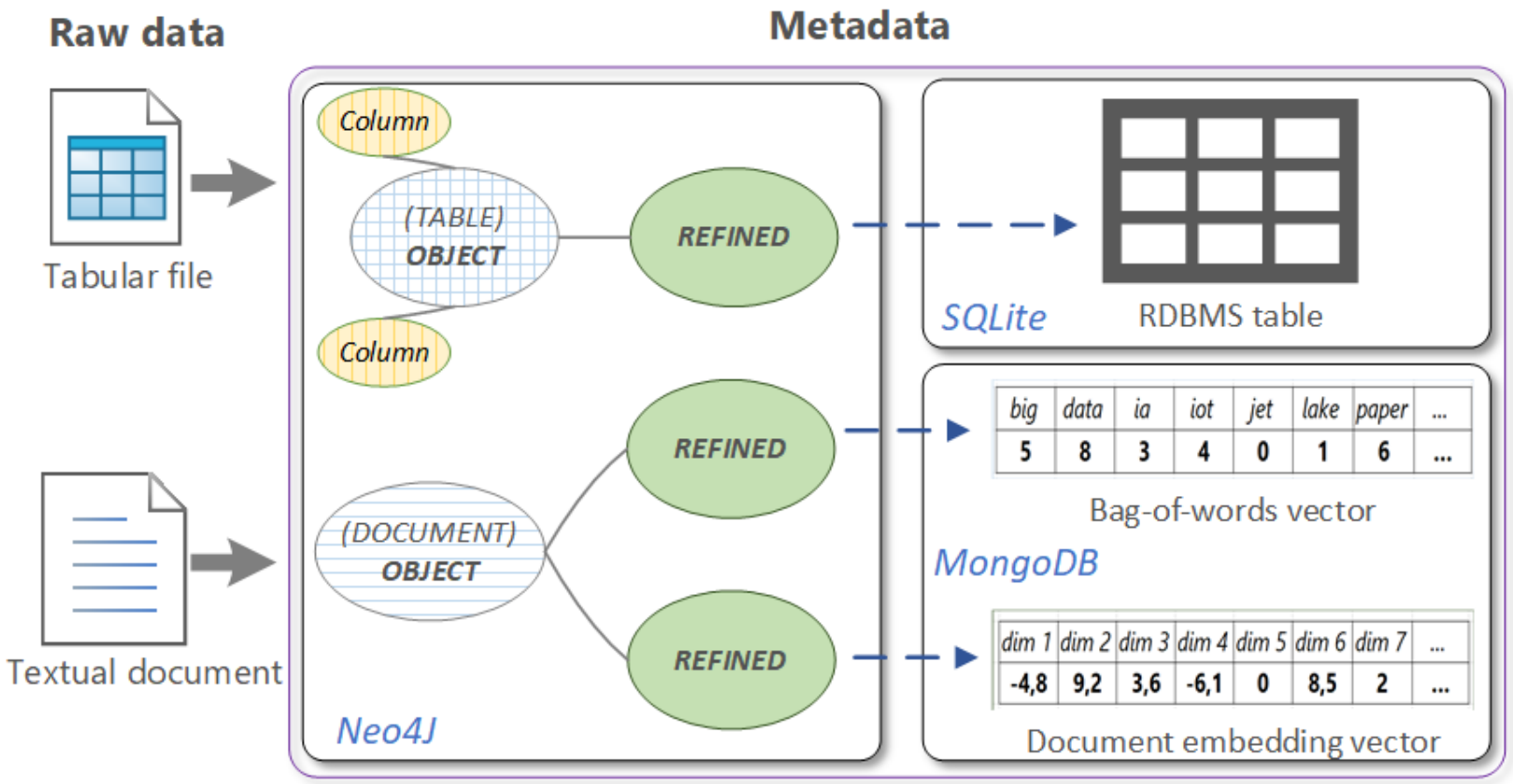
Réseau de recherche *datalakers*

Exemple : AUDAL

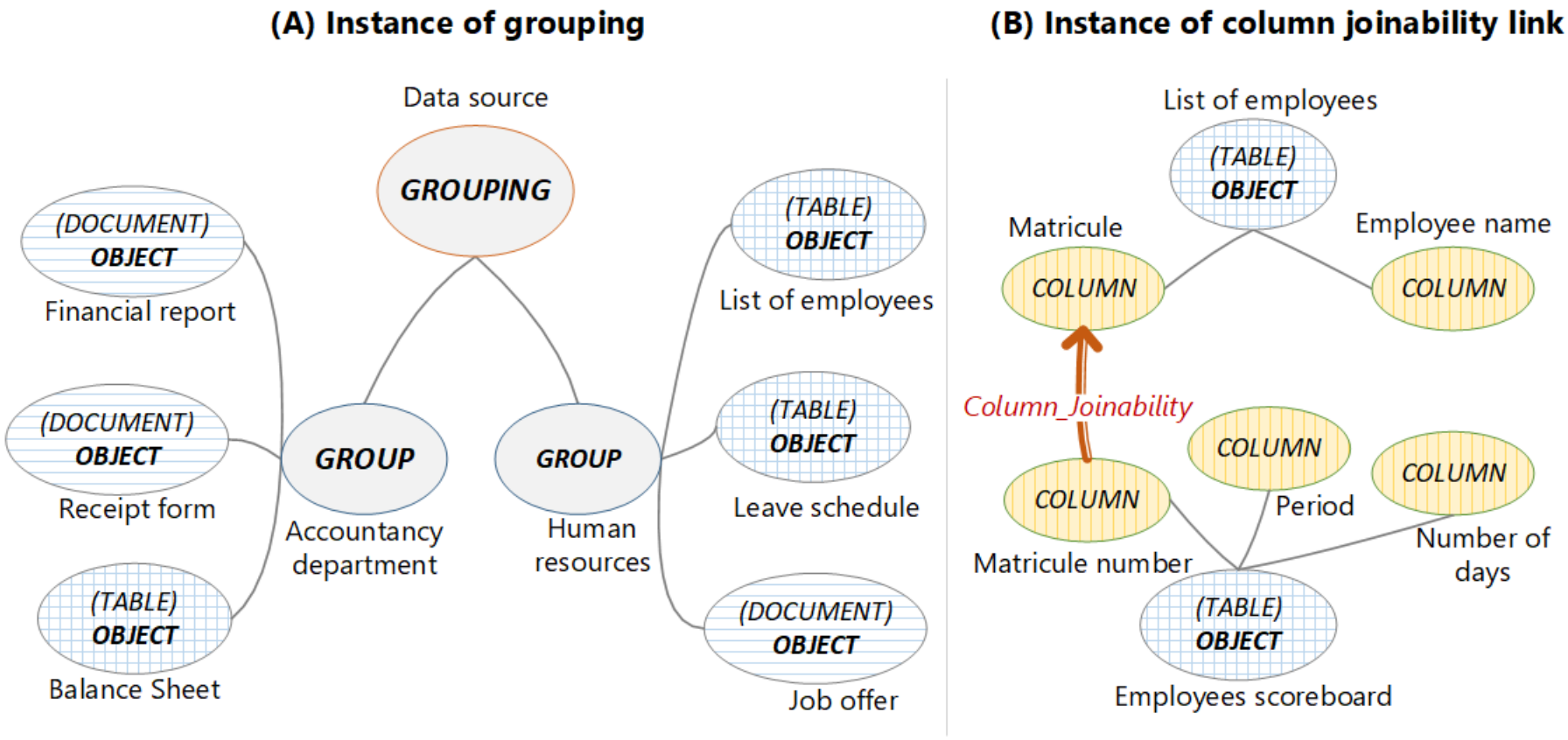
Sawadogo et Darmont 2021

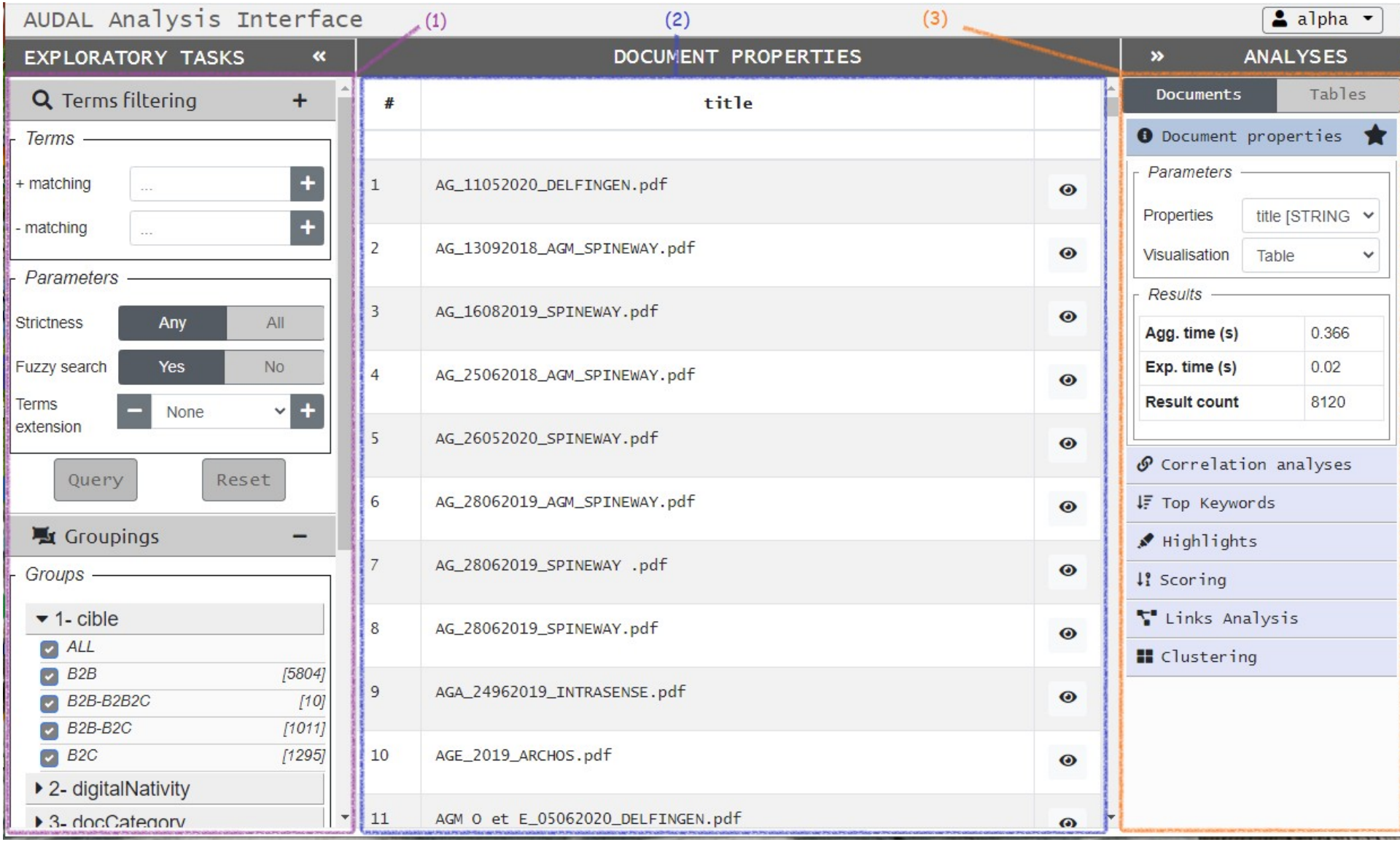


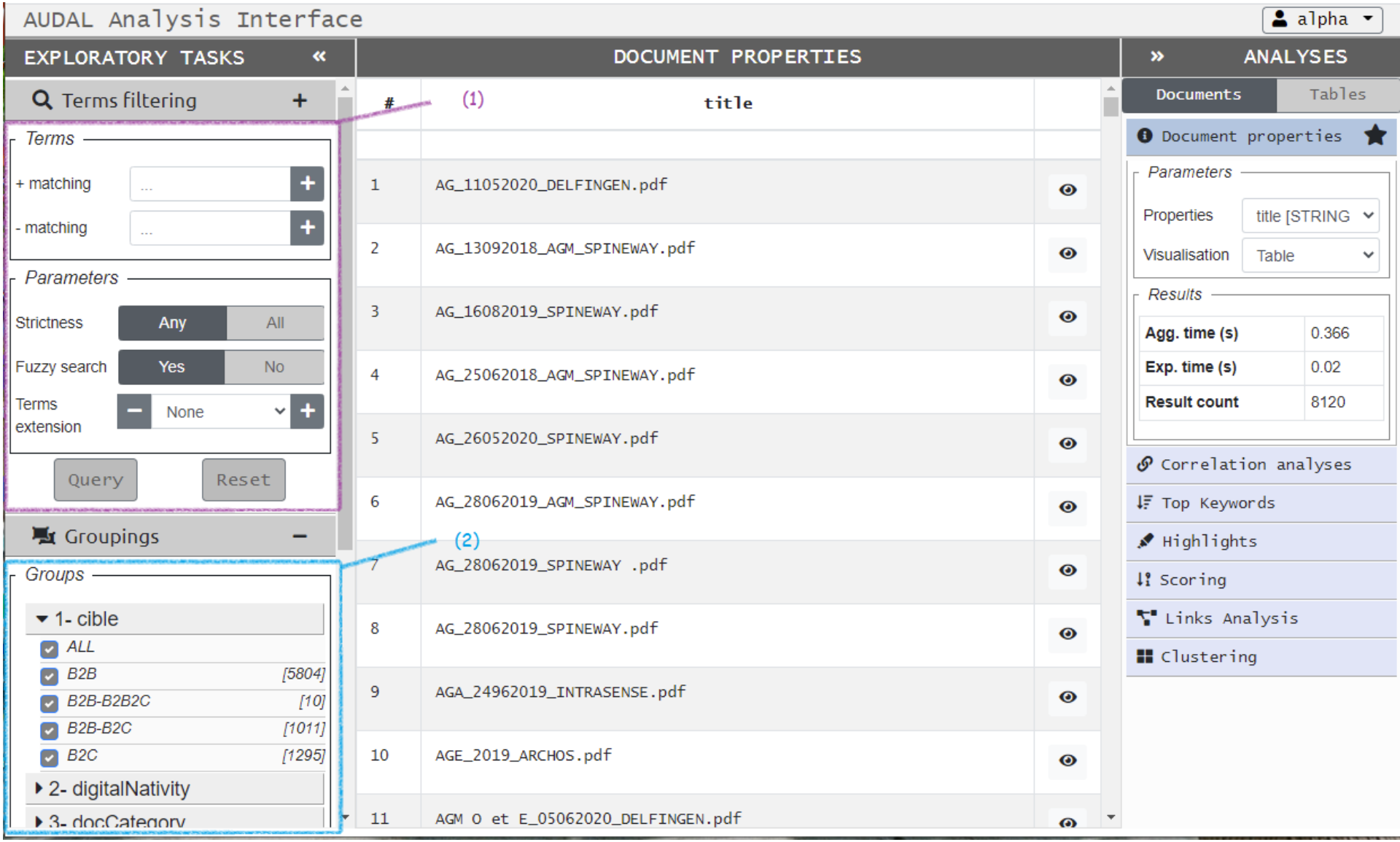
Métadonnées intra-objet d'AUDAL

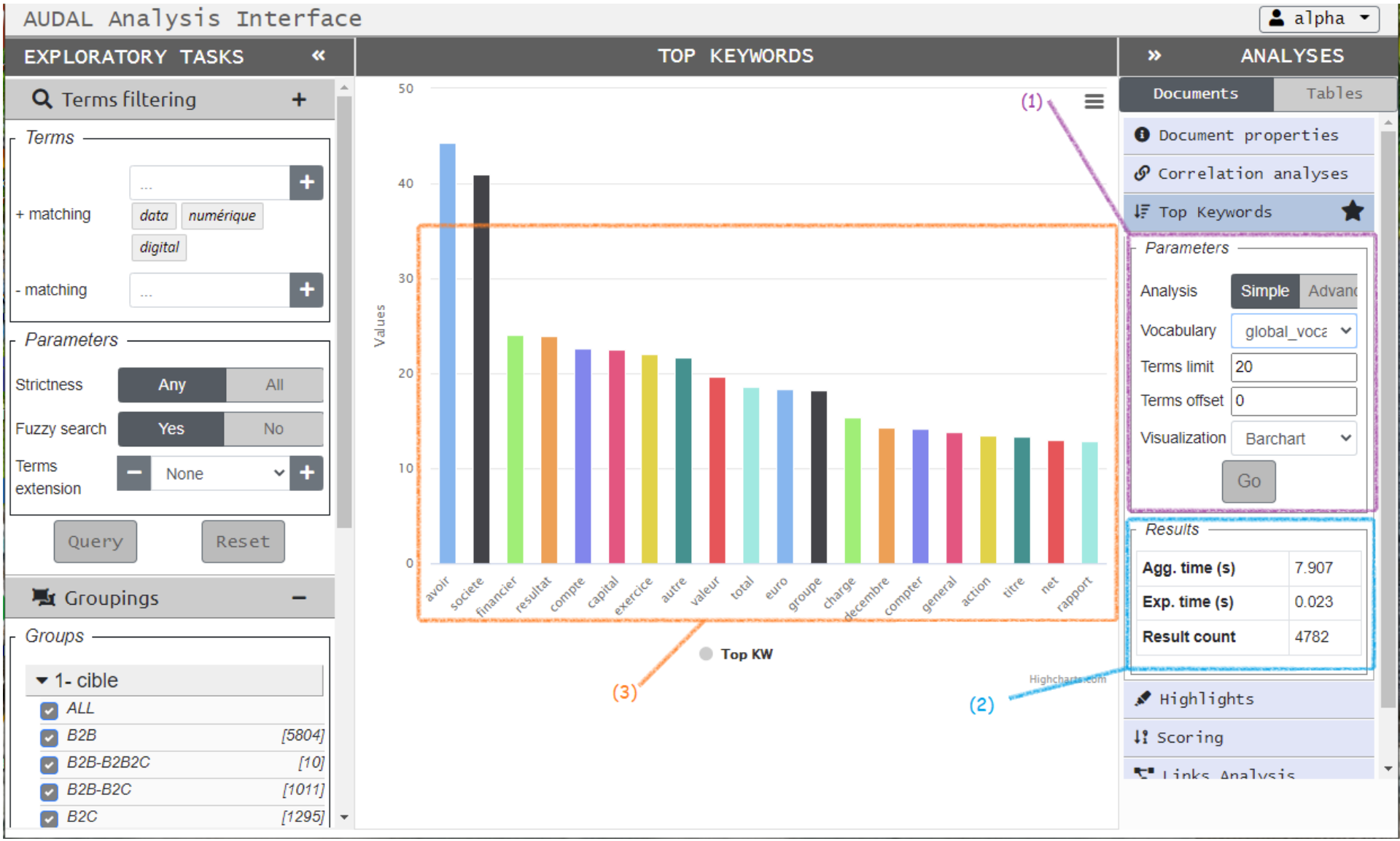


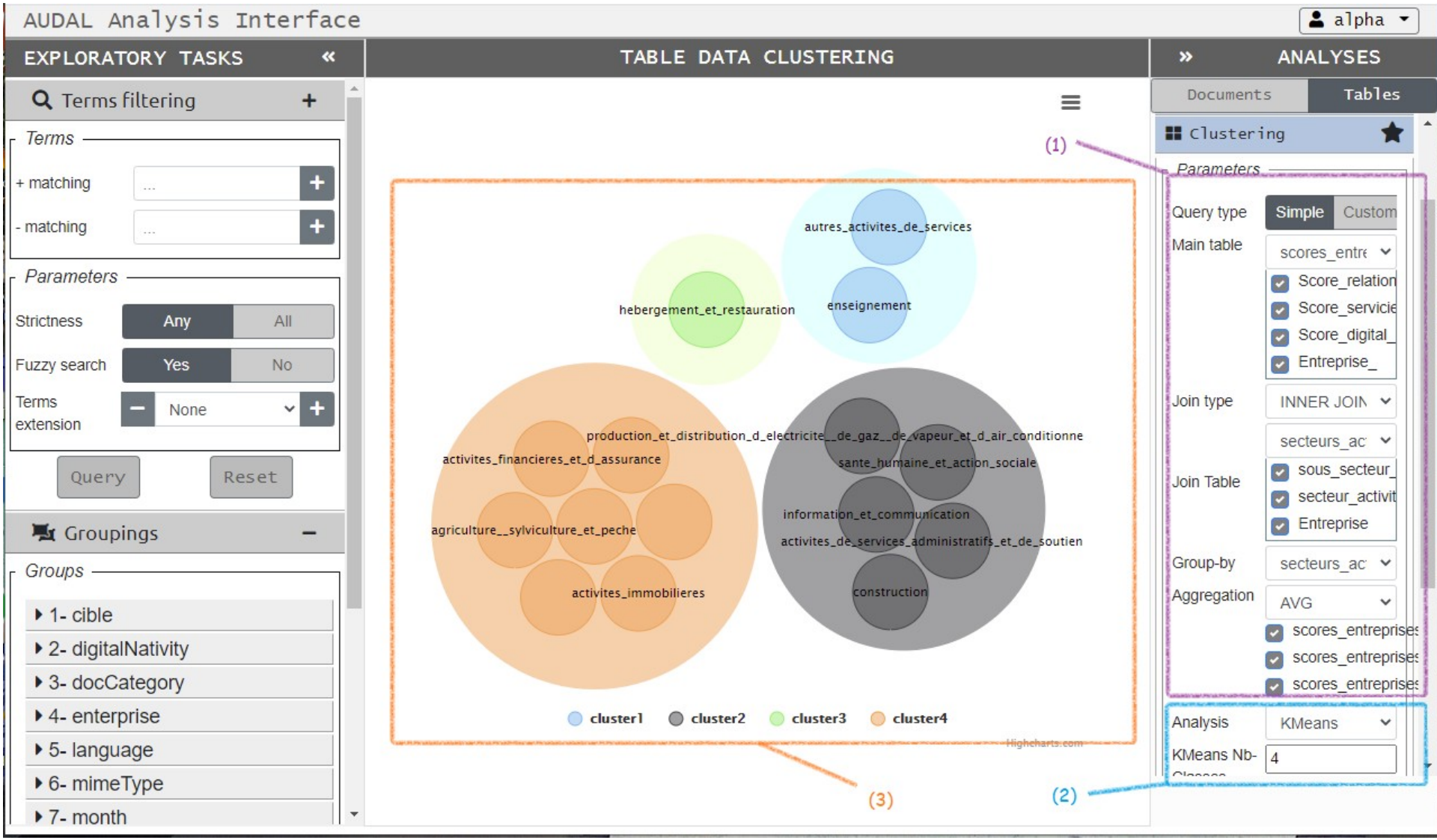
Métadonnées inter-objets d'AUDAL




AUDAL Analysis Interface
(1)
(2)
(3)
alpha
EXPLORATORY TASKS
DOCUMENT PROPERTIES
ANALYSES
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]
2- digitalNativity
3- docCategory
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 O et E_05062020_DELFINGEN.pdf
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


AUDAL Analysis Interface
alpha
EXPLORATORY TASKS
DOCUMENT PROPERTIES
ANALYSES
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]
2- digitalNativity
3- docCategory
(1)
(2)
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 O et E_05062020_DELFINGEN.pdf
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


AUDAL Analysis Interface
alpha
EXPLORATORY TASKS
TOP KEYWORDS
ANALYSES
Terms filtering
Terms
+ matching
data
numérique
digital
- 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]
Values
50
40
30
20
10
0
avoir
societe
financier
resultat
compte
capital
exercice
autre
valeur
total
euro
groupe
charge
decembre
compter
general
action
titre
net
rapport
Top KW
Highcharts.com
(1)
(2)
(3)
Documents
Tables
Document properties
Correlation analyses
Top Keywords
Parameters
Analysis
Simple
Advanc
Vocabulary
global_voca
Terms limit
20
Terms offset
0
Visualization
Barchart
Go
Results
Agg. time (s) 7.907
Exp. time (s) 0.023
Result count 4782
Highlights
Scoring
Links Analysis


AUDAL Analysis Interface
alpha
EXPLORATORY TASKS
TABLE DATA CLUSTERING
ANALYSES
Terms filtering
Terms
+ matching
- matching
Parameters
Strictness
Any
All
Fuzzy search
Yes
No
Terms extension
None
Query
Reset
Groupings
Groups
1- cible
2- digitalNativity
3- docCategory
4- enterprise
5- language
6- mimeType
7- month
autres_activites_de_services
hebergement_et_restauration
enseignement
production_et_distribution_d_electricite__de_gaz__de_vapeur_et_d_air_conditionne
activites_financieres_et_d_assurance
sante_humaine_et_action_sociale
information_et_communication
agriculture__sylviculture_et_peche
activites_de_services_administratifs_et_de_soutien
activites_immobilieres
construction
cluster1
cluster2
cluster3
cluster4
Highcharts.com
(1)
(2)
(3)
Documents
Tables
Clustering
Parameters
Query type
Simple
Custom
Main table
scores_entre
Score_relation
Score_servicie
Score_digital_
Entreprise_
Join type
INNER JOIN
secteurs_ac
Join Table
sous_secteur_
secteur_activit
Entreprise
Group-by
secteurs_ac
Aggregation
AVG
scores_entreprises
scores_entreprises
scores_entreprises
Analysis
KMeans
KMeans Nb-
4

Plan

- ✓ Définitions
- ✓ Entrepôts et lacs d
- ✓ Métadonnées et étadonnées
- ✓ Architectures et te pour les lacs
- ✓ Discussion, travaux de recherche