DATA LAKES: A NEW PARADIGM FOR DATA PLATFORMS AND CURRENT CHALLENGES

Anne Laurent

LIRMM
POLYTECH MONTPELLIER
ISDM INSTITUT DE SCIENCE DES DONNÉES MONTPELLIER
MESO@LR
UNIVERSITÉ DE MONTPELLIER

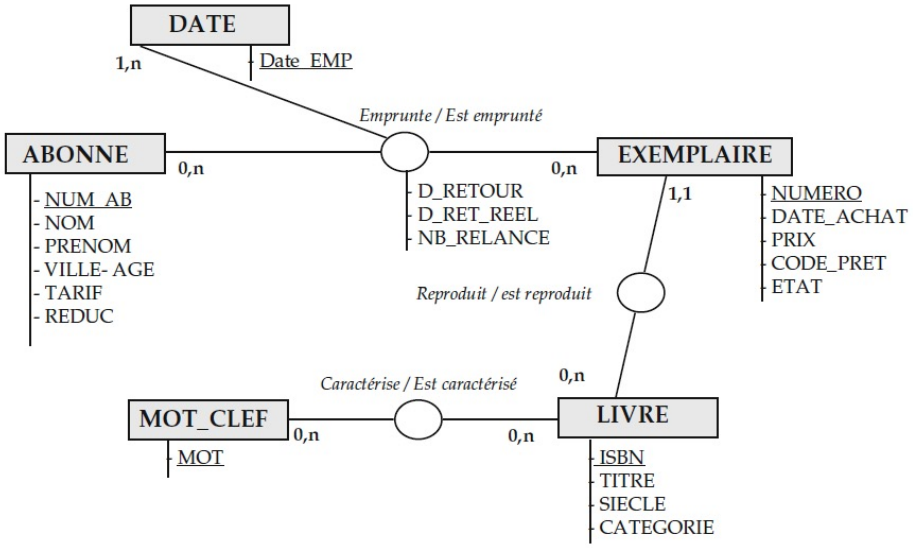
From « information-driven » architectures to « data-driven » architectures

- Multiplication of data sources « without any a priori usage », as for example IoT
- Augmentation of semi-structured sources
- Available technical solutions

→ **Data lakes**: first seen as commercial tools to deal with large amounts of data and then coming to a new paradigm

Operational Systems

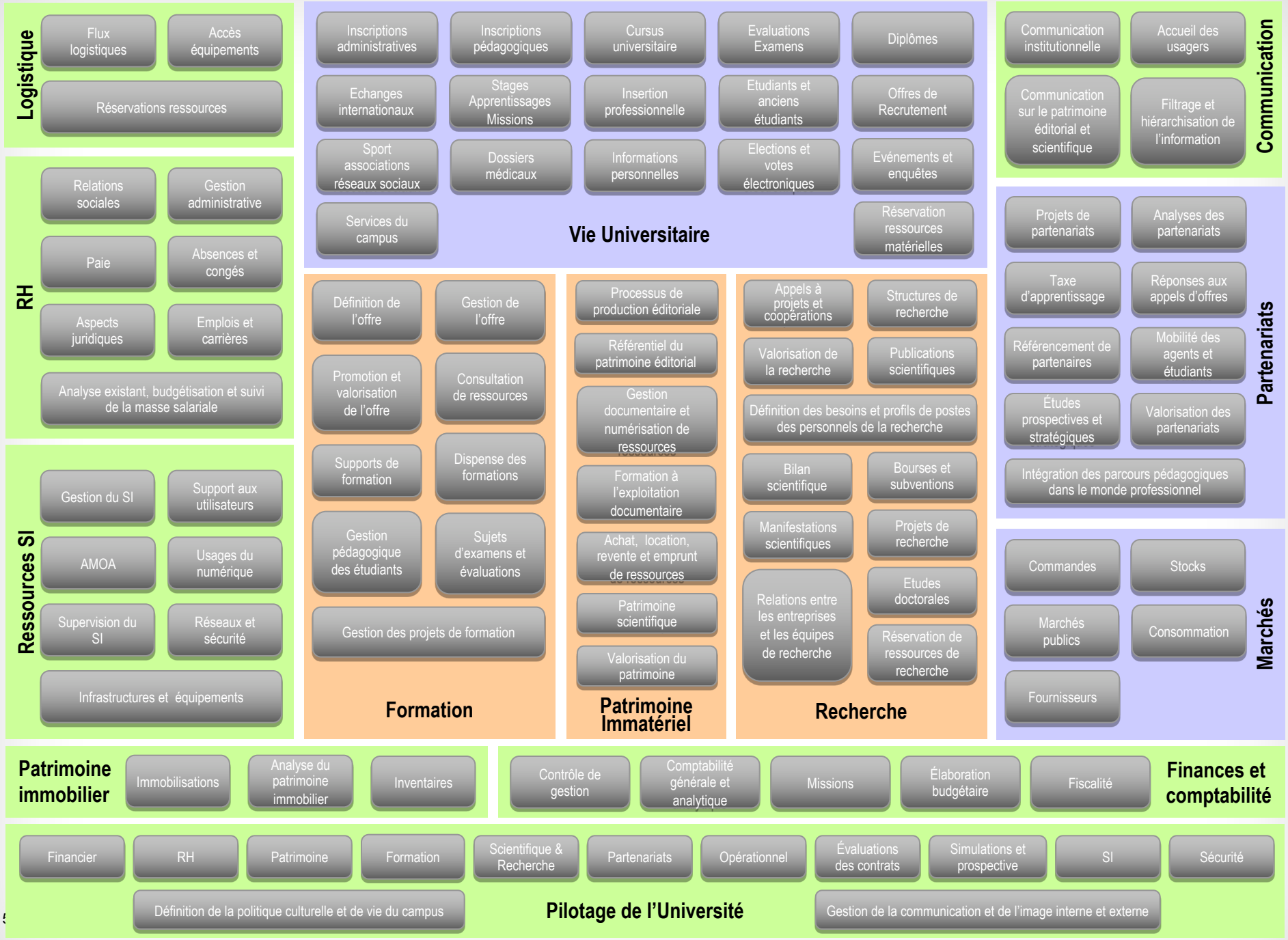
- Automatisation
- From the early 70's (Codd)

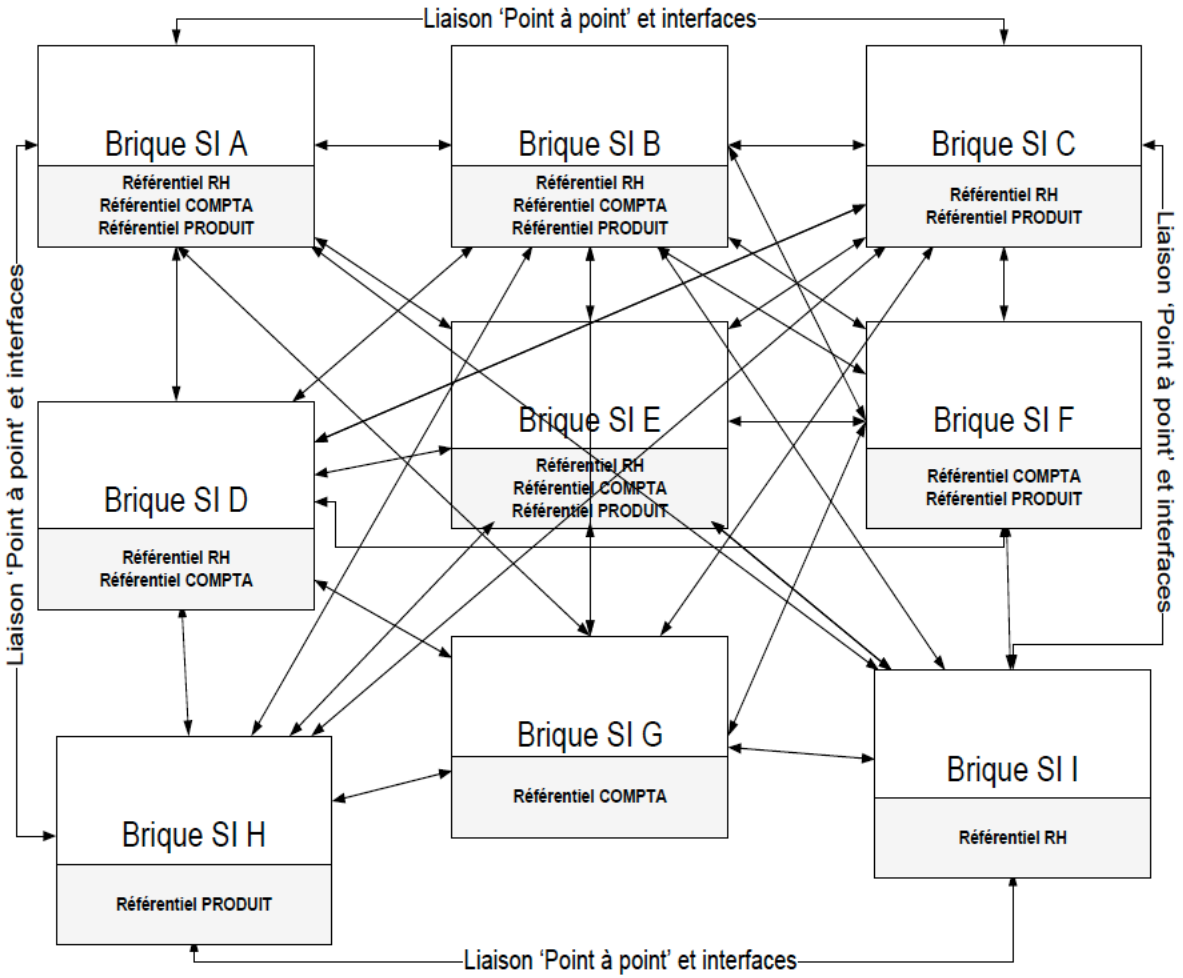


Challenges

- Consistency
- Transactional
- Scalability
- Master Data
- Agility

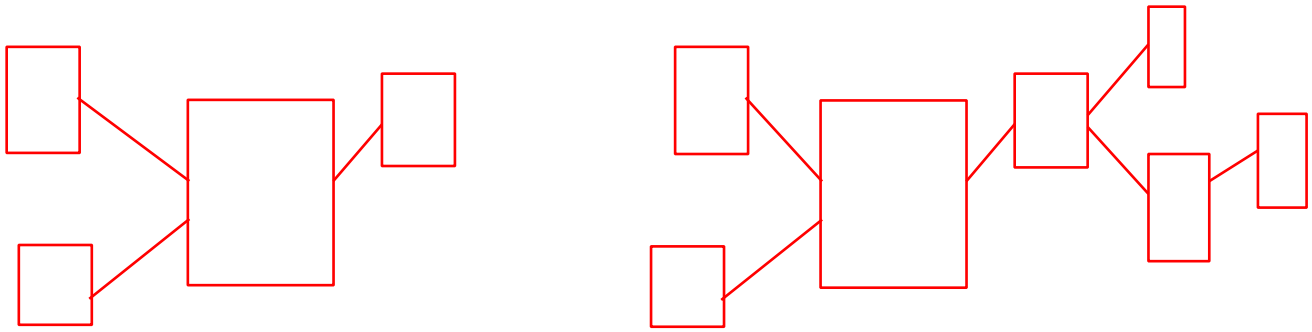
→ More an more complex architectures


Logistique
Flux logistiques
Accès équipements
Réservations ressources
RH
Relations sociales
Gestion administrative
Paie
Absences et congés
Aspects juridiques
Emplois et carrières
Analyse existant, budgétisation et suivi de la masse salariale
Ressources SI
Gestion du SI
Support aux utilisateurs
AMOA
Usages du numérique
Supervision du SI
Réseaux et sécurité
Infrastructures et équipements
Inscriptions administratives
Inscriptions pédagogiques
Cursus universitaire
Evaluations Examens
Diplômes
Echanges internationaux
Stages Apprentissages Missions
Insertion professionnelle
Etudiants et anciens étudiants
Offres de Recrutement
Sport associations réseaux sociaux
Dossiers médicaux
Informations personnelles
Elections et votes électroniques
Evénements et enquêtes
Services du campus
Réservation ressources matérielles
Vie Universitaire
Communication institutionnelle
Accueil des usagers
Communication sur le patrimoine éditorial et scientifique
Filtrage et hiérarchisation de l'information
Communication
Définition de l'offre
Gestion de l'offre
Promotion et valorisation de l'offre
Consultation de ressources
Supports de formation
Dispense des formations
Gestion pédagogique des étudiants
Sujets d'examens et évaluations
Gestion des projets de formation
Formation
Processus de production éditoriale
Référentiel du patrimoine éditorial
Gestion documentaire et numérisation de ressources
Formation à l'exploitation documentaire
Achat, location, revente et emprunt de ressources
Patrimoine scientifique
Valorisation du patrimoine
Patrimoine Immatériel
Appels à projets et coopérations
Structures de recherche
Valorisation de la recherche
Publications scientifiques
Définition des besoins et profils de postes des personnels de la recherche
Bilan scientifique
Bourses et subventions
Manifestations scientifiques
Projets de recherche
Relations entre les entreprises et les équipes de recherche
Etudes doctorales
Réservation de ressources de recherche
Recherche
Projets de partenariats
Analyses des partenariats
Taxe d'apprentissage
Réponses aux appels d'offres
Référencement de partenaires
Mobilité des agents et étudiants
Études prospectives et stratégiques
Valorisation des partenariats
Intégration des parcours pédagogiques dans le monde professionnel
Partenariats
Commandes
Stocks
Marchés publics
Consommation
Fournisseurs
Marchés
Patrimoine immobilier
Immobilisations
Analyse du patrimoine immobilier
Inventaires
Contrôle de gestion
Comptabilité générale et analytique
Missions
Élaboration budgétaire
Fiscalité
Finances et comptabilité
Financier
RH
Patrimoine
Formation
Scientifique & Recherche
Partenariats
Opérationnel
Évaluations des contrats
Simulations et prospective
SI
Sécurité
Définition de la politique culturelle et de vie du campus
Pilotage de l'Université
Gestion de la communication et de l'image interne et externe


Liaison 'Point à point' et interfaces
Brique SI A
Référentiel RH
Référentiel COMPTA
Référentiel PRODUIT
Brique SI B
Référentiel RH
Référentiel COMPTA
Référentiel PRODUIT
Brique SI C
Référentiel RH
Référentiel PRODUIT
Liaison 'Point à point' et interfaces
Liaison 'Point à point' et interfaces
Brique SI D
Référentiel RH
Référentiel COMPTA
Brique SI E
Référentiel RH
Référentiel COMPTA
Référentiel PRODUIT
Brique SI F
Référentiel COMPTA
Référentiel PRODUIT
Brique SI G
Référentiel COMPTA
Brique SI H
Référentiel PRODUIT
Brique SI I
Référentiel RH
Liaison 'Point à point' et interfaces

Decisional Information Systems

- Measures/KPI, dimensions, ETL, data warehouses/marts - 90's (Codd)
- BCNF, 3NF,…



Evolution of architectures

- Towards schema « on read » models versus « schema « on write » - ETL/ELT
- No Data Silos

*"If you think of a Data Mart 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 can come to examine, dive in, or take samples."*
Dixon, 2010

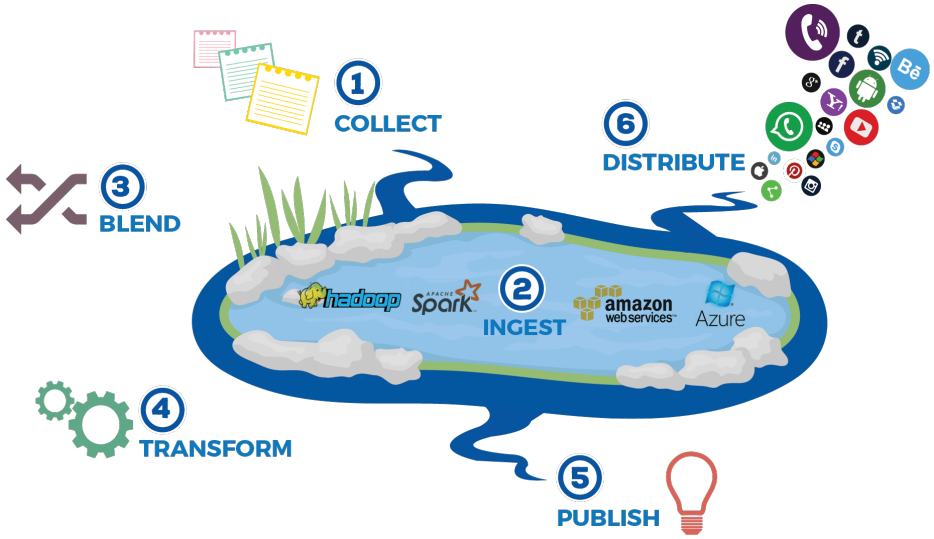
Data Lakes (C. Madera)

A data lake is a collection of data with the following principles:

- No predefined schema,
- All data formats accepted,
- Raw data,
- Conceptually seen as a unique storage but not necessary materialized,
- Accessed by data scientists,
- Provided with a data catalog of meta-data,
- Provided with data governance rules.

"A "Data Lake" is a methodology enabled by a massive data repository based on low cost technologies that improves the capture, refinement, archival, and exploration of raw data within an enterprise. "

Fang, 2015

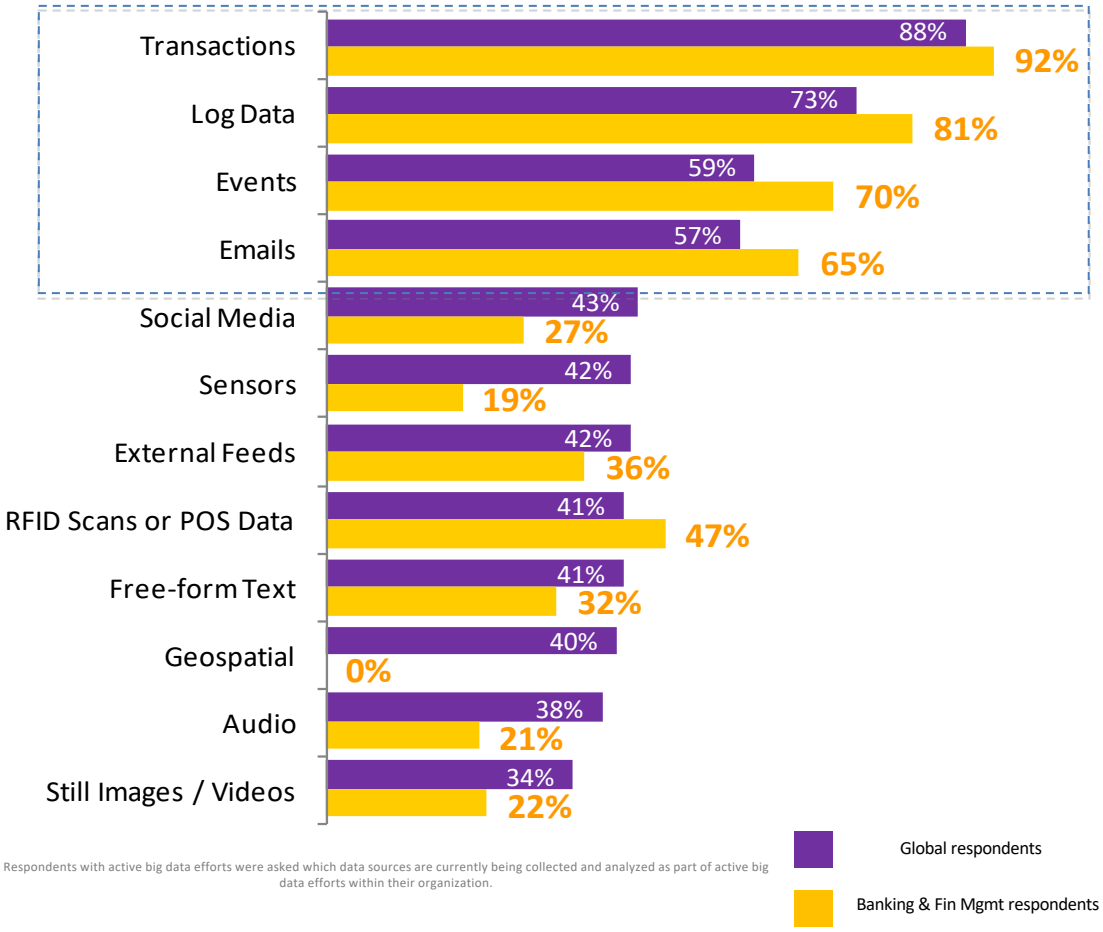


https://www.gekko.fr/batir-un-datalake-sur-aws-quelles-solutions/

WHAT IS FEEDING THE LAKE?

- 80% business data
  - From the information system
  - From the data warehouses
  - From tweets and messages
  - From softwares and web logs
  - From digital assets
- External - Open Data

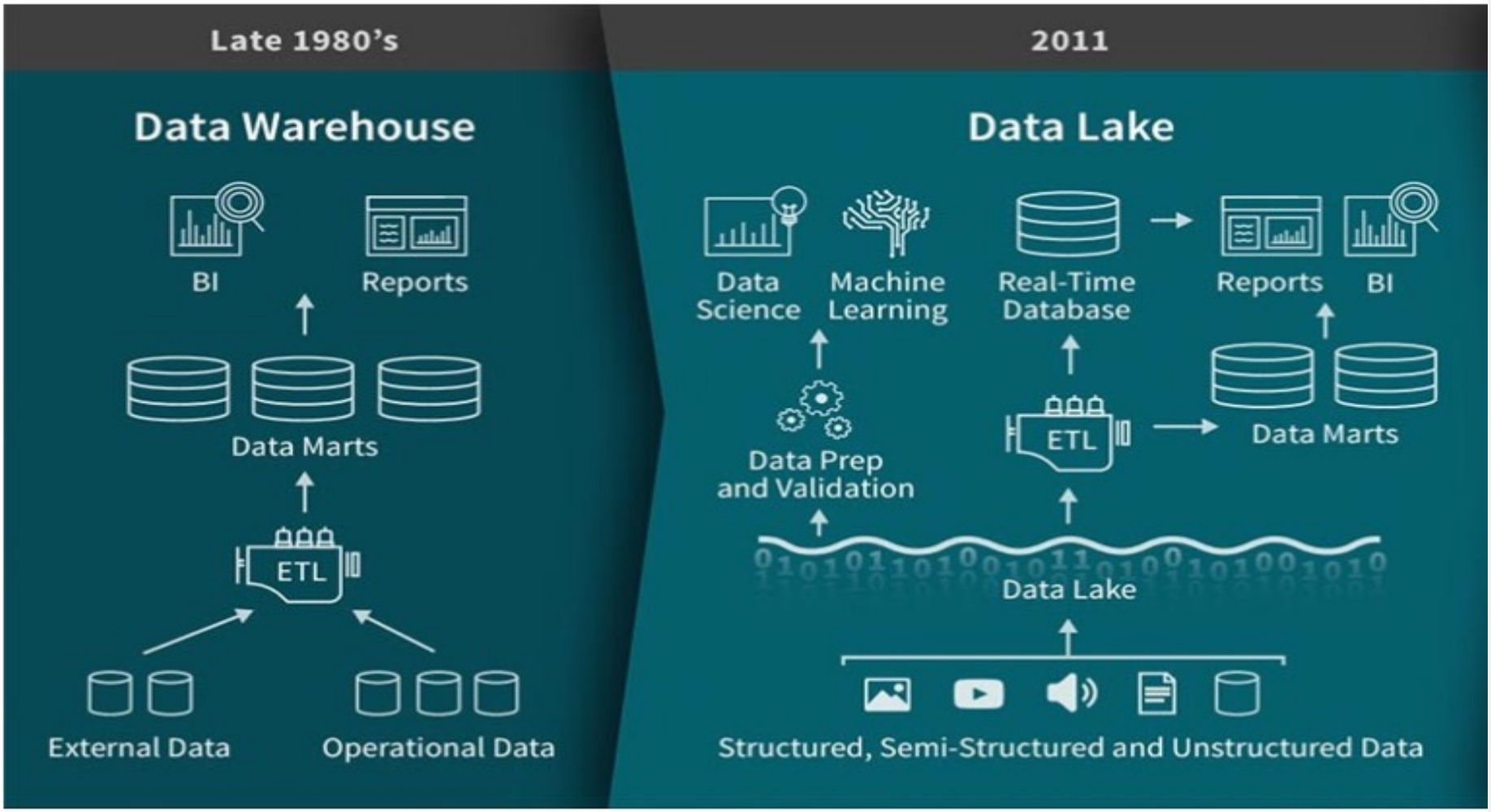
Big data sources



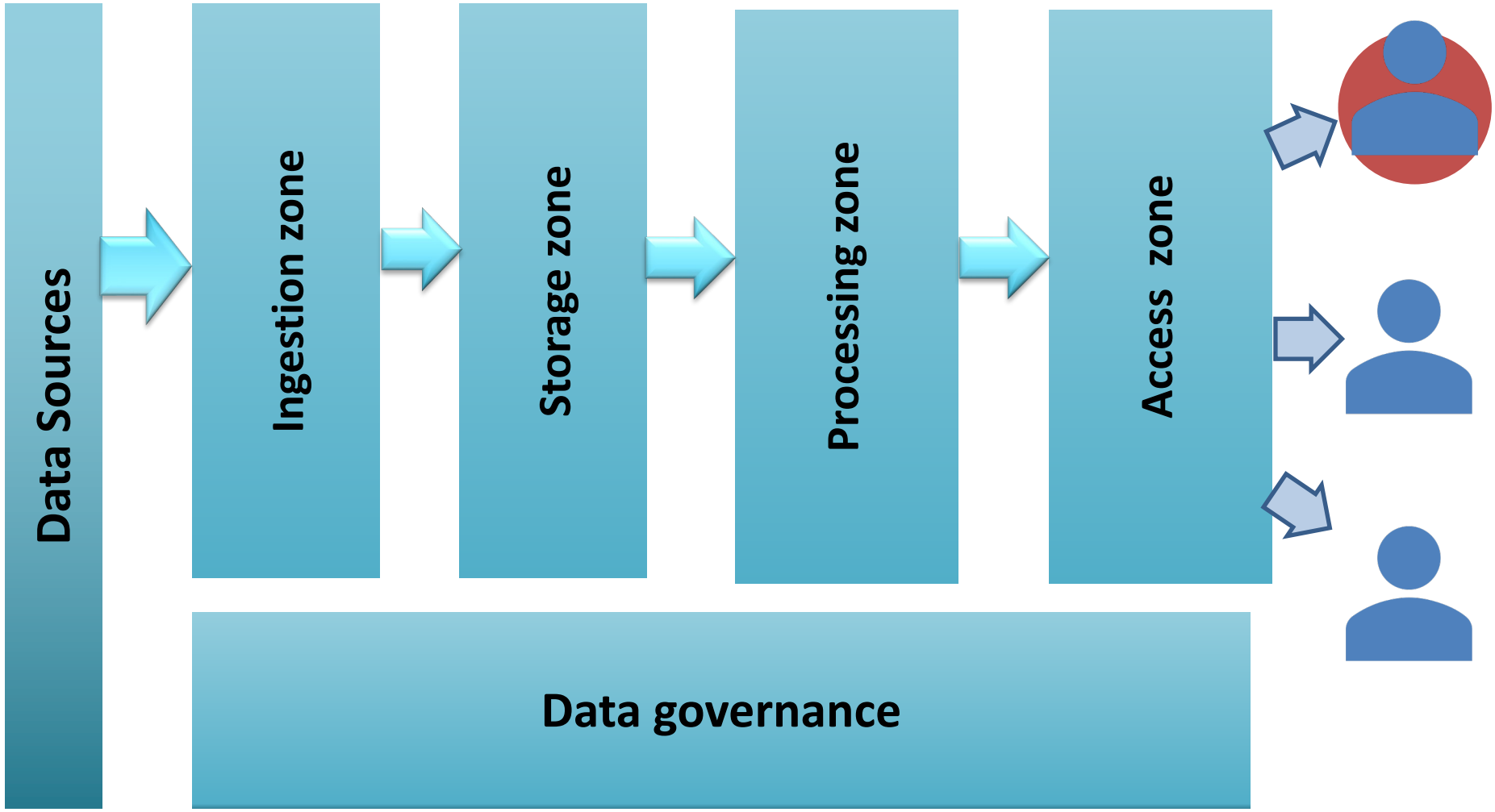
Source: The real world use of Big Data, IBM & University of Oxford

schema « on write »

schema « on read »



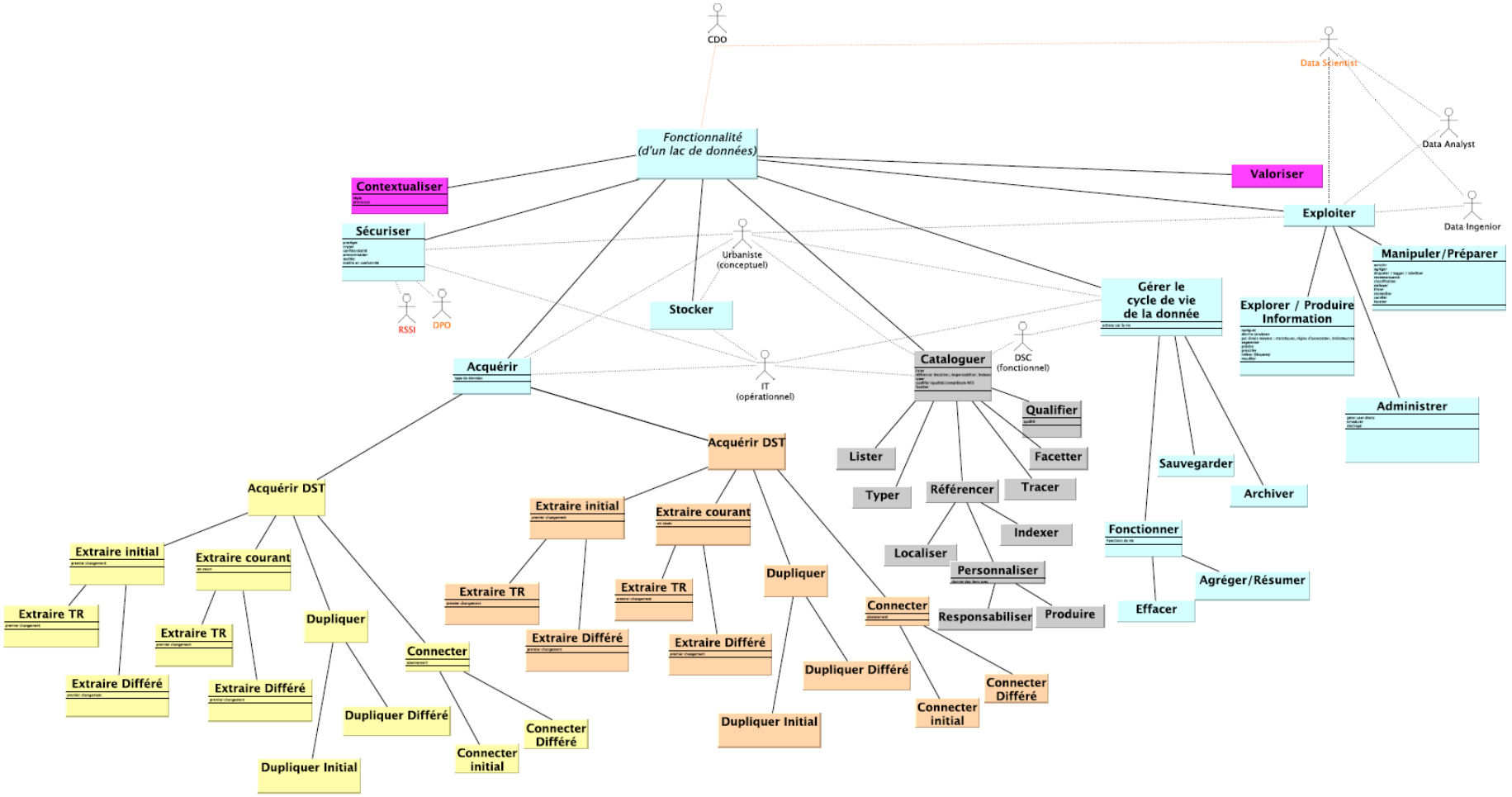
@Databricks


Data Sources
Ingestion zone
Storage zone
Processing zone
Access zone
Data governance

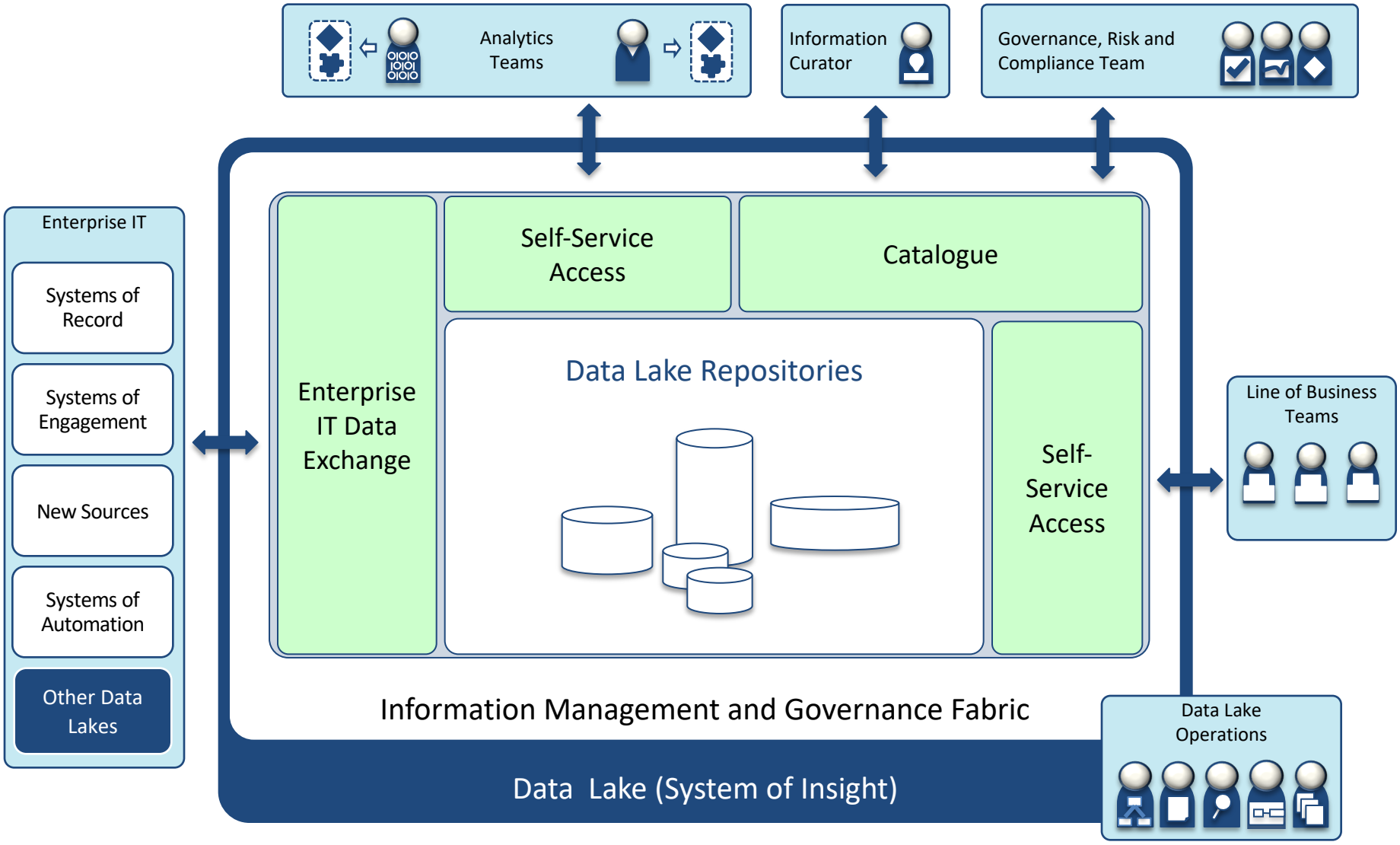
Defining the data lake architectures

Main features:

- Collect
- Catalog
- Manage data life cycle
- Exploit
- Secure
- Store

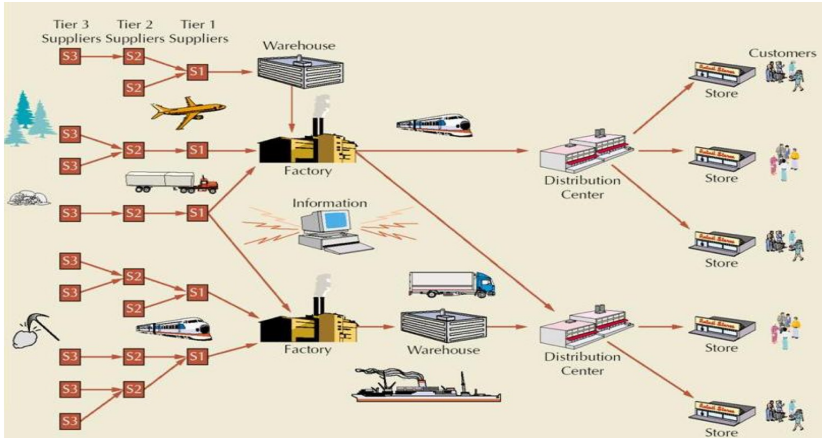


C. Madera, M. Huchard, A. Laurent, A. Miralles


Analytics Teams
Information Curator
Governance, Risk and Compliance Team
Enterprise IT
Systems of Record
Systems of Engagement
New Sources
Systems of Automation
Other Data Lakes
Enterprise IT Data Exchange
Self-Service Access
Catalogue
Data Lake Repositories
Self-Service Access
Line of Business Teams
Information Management and Governance Fabric
Data Lake Operations
Data Lake (System of Insight)

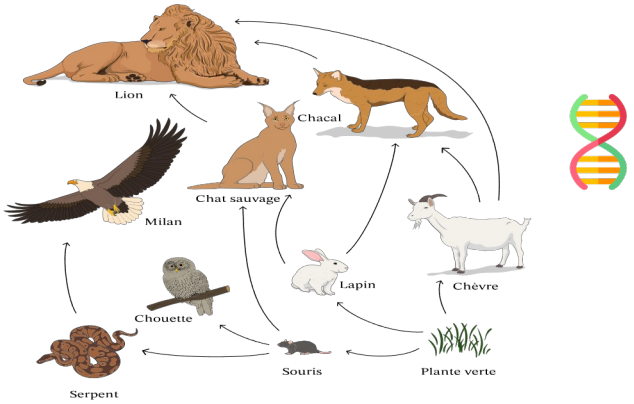
ANALOGIES

Supply Chain



Ref:https://supernet.isenberg.umass.edu/courses/OIM413-Fall2017/OIM413lecture2.pdf

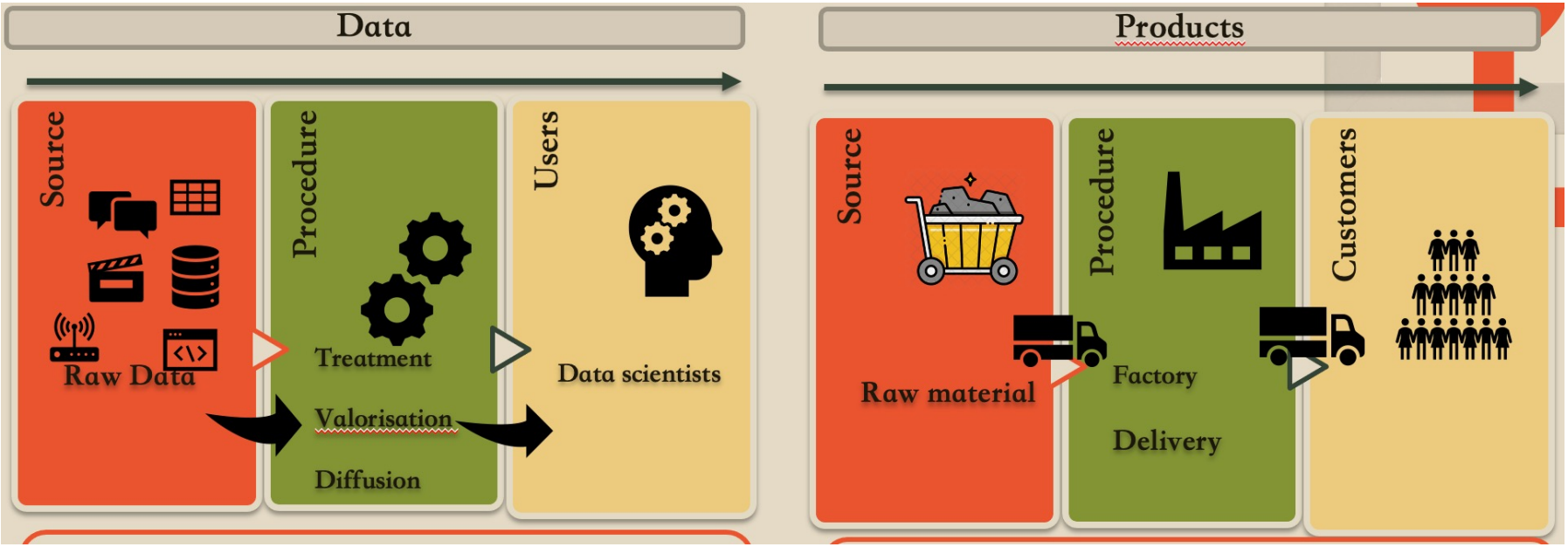
Biology



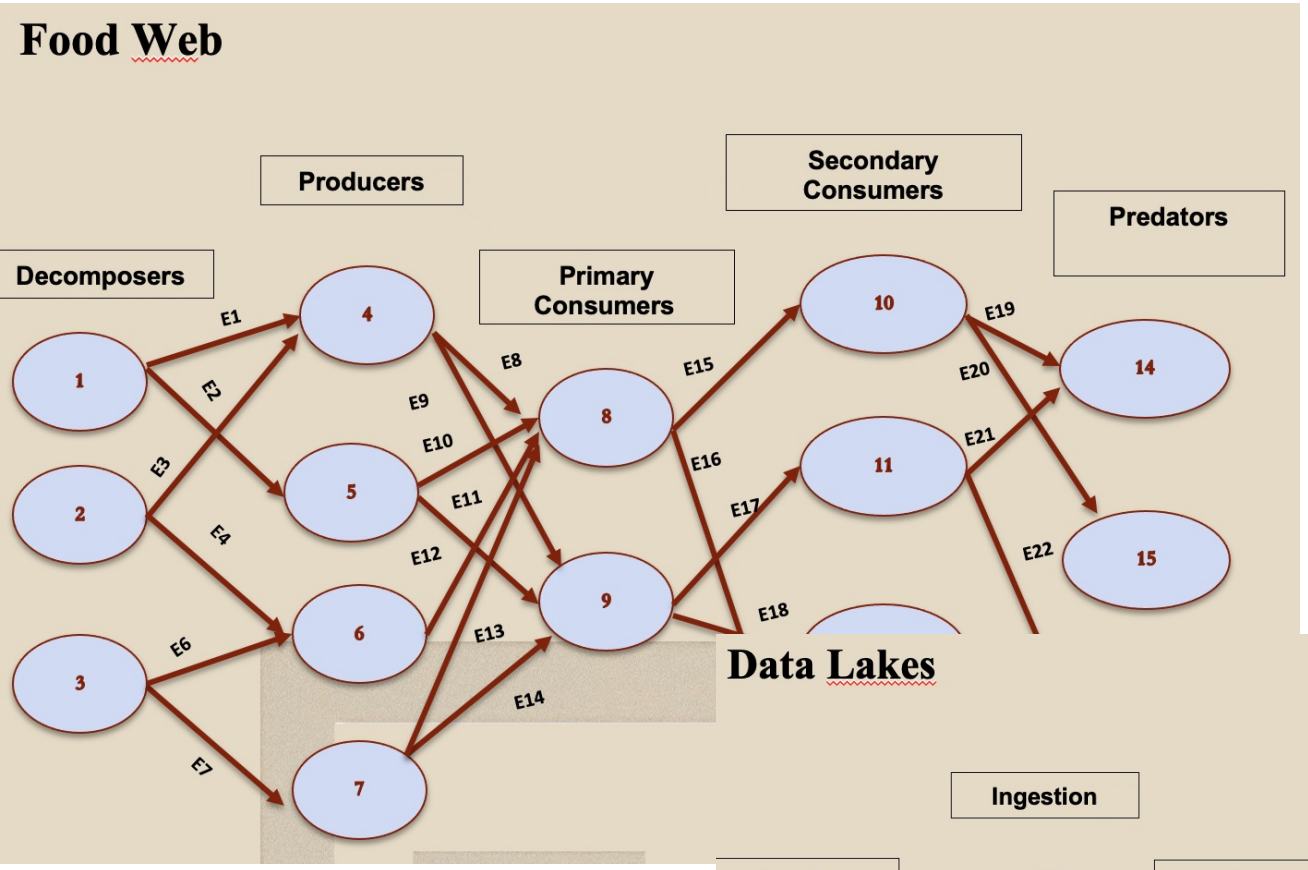
Ref:https://www.nagwa.com/fr/worksheets/295181509468/

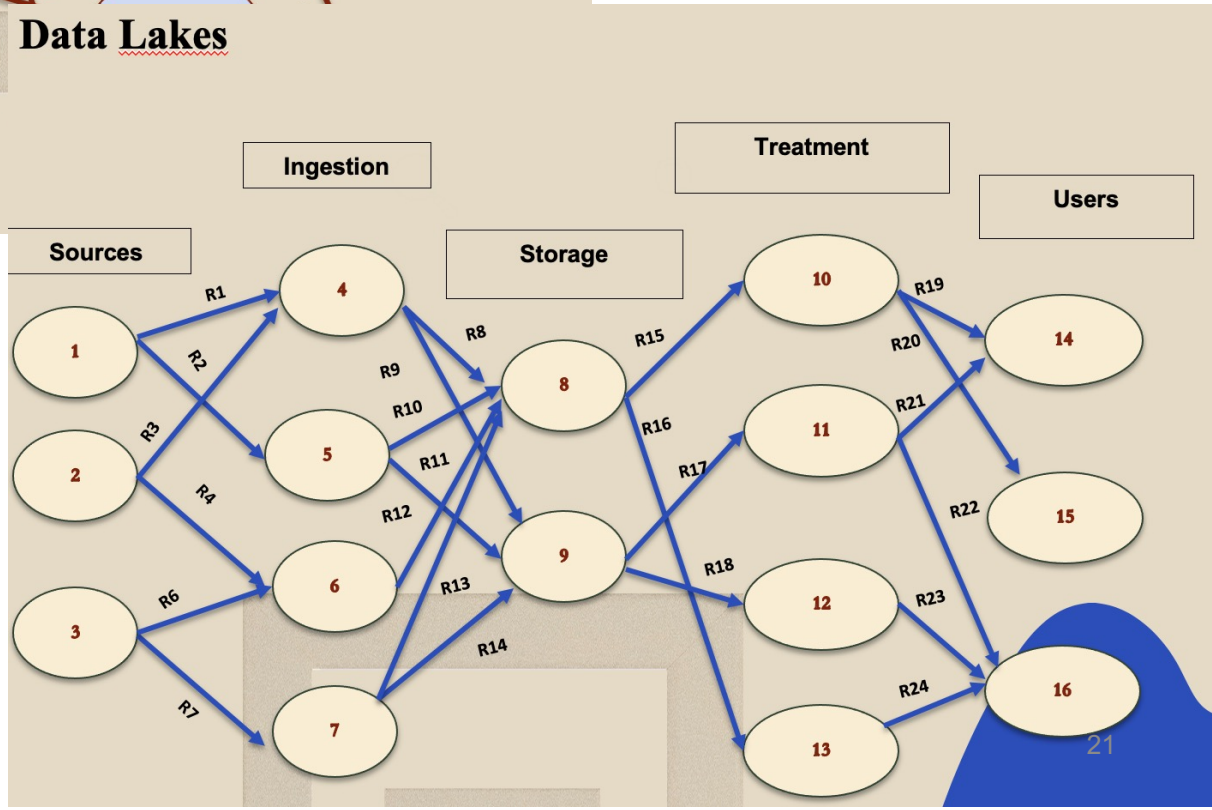
Module	Supply Chain	Natural Lake (Species and Ecosystems)	Data Lake
Members/Levels	Supplier Manufacturer Distributor Retailer Customer	Ecosystem Components (Animals, Plants, Microorganisms) Biological processes (Breed, Birth, Growth and Death) Ecological processes (Eat and be Eaten)	Ingestion stage Storage stage Processing stage Access stage Final user
Products	Commodity (Forward Flow) Information (Backward Flow)	Biodiversity (Species diversity) Ecological complexity (More Species More complex) Biomass	Data
Management Strategies	Lean SCM Agile SCM Postponement SCM Speculation SCM Green Supply Chain	Species evolution (Mutation, Recombination, Drift, Selection) Competition Parasitism (Negative association) Mutualism (Positive association) Predation	Metadata management Data management Data Governance
Objective Functions	Cost minimization Sales maximization Profit maximization Lead time minimization	At species level: Maximize reproduction and survive (Fitness) At ecosystem level: Maximize resilience	Cost minimization Fill rate maximization Response time minimization

Derakhshannia et al.



Supply Chain

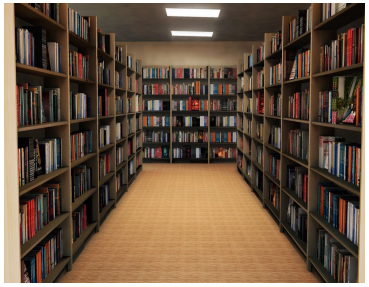

Food Web
Producers
Secondary Consumers
Predators
Decomposers
Primary Consumers
1
2
3
4
5
6
7
8
9
10
11
14
15
E1
E2
E3
E4
E6
E7
E8
E9
E10
E11
E12
E13
E14
E15
E16
E17
E18
E19
E20
E21
E22


Data Lakes
Ingestion
Treatment
Users
Sources
Storage
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
R1
R2
R3
R4
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24

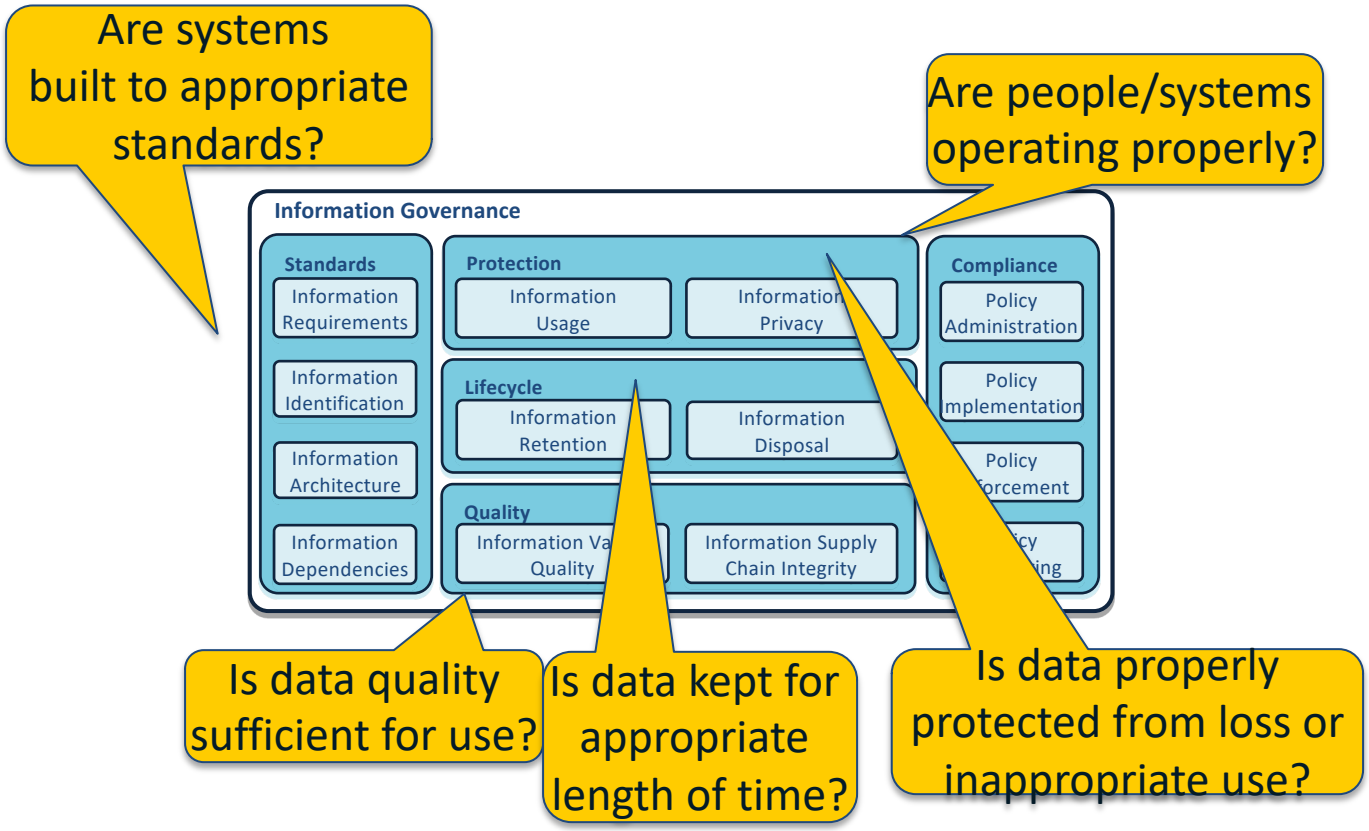
Biology

LAKES AND SWAMPS

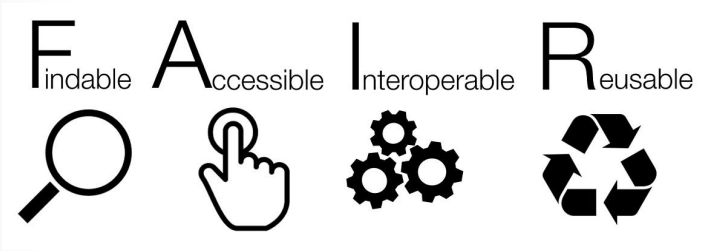
DATA GOVERNANCE



INFORMATION GOVERNANCE



FAIR PRINCIPLES



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

THE KEY CONCEPT OF META-DATA

Date	09/2015					
Autor	John Doe					
Label: Label1						
Mode				Measurement from above		
Emission wavelength start				380	nm	
Emission wavelength end				600	nm	
Emissions wavelength step				2	nm	
Scan count				111		
Spectrum (Em)				280...850: 20 nm		
Spectrum (ex) (Sector 1)				230...315: 5 nm		
Spectrum (ex) (Sector 2)				316...850: 10 nm		
	Temperature: 25.5 °C					
WL	380	382	384	386	388	390
E1	966	224	162	171	206	273
E2	477	240	135	168	148	150
E3	627	235	171	174	232	263
E4	280	160	147	214	252	375
E5	657	245	164	167	157	179
E6	159	97	95	101	150	171

Fig. 1. Spreadsheet Example

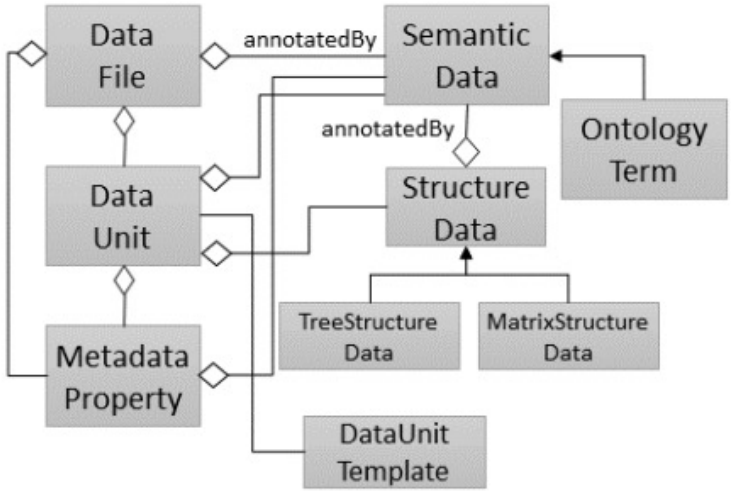


Fig. 2. Conceptual View of the Model

C. Quix et al.

METADATA MODELS

Oram, 2015 –

three distinct types of metadata, i.e., Business Metadata, Operational Metadata, and Technical Metadata

C. Quix et al.

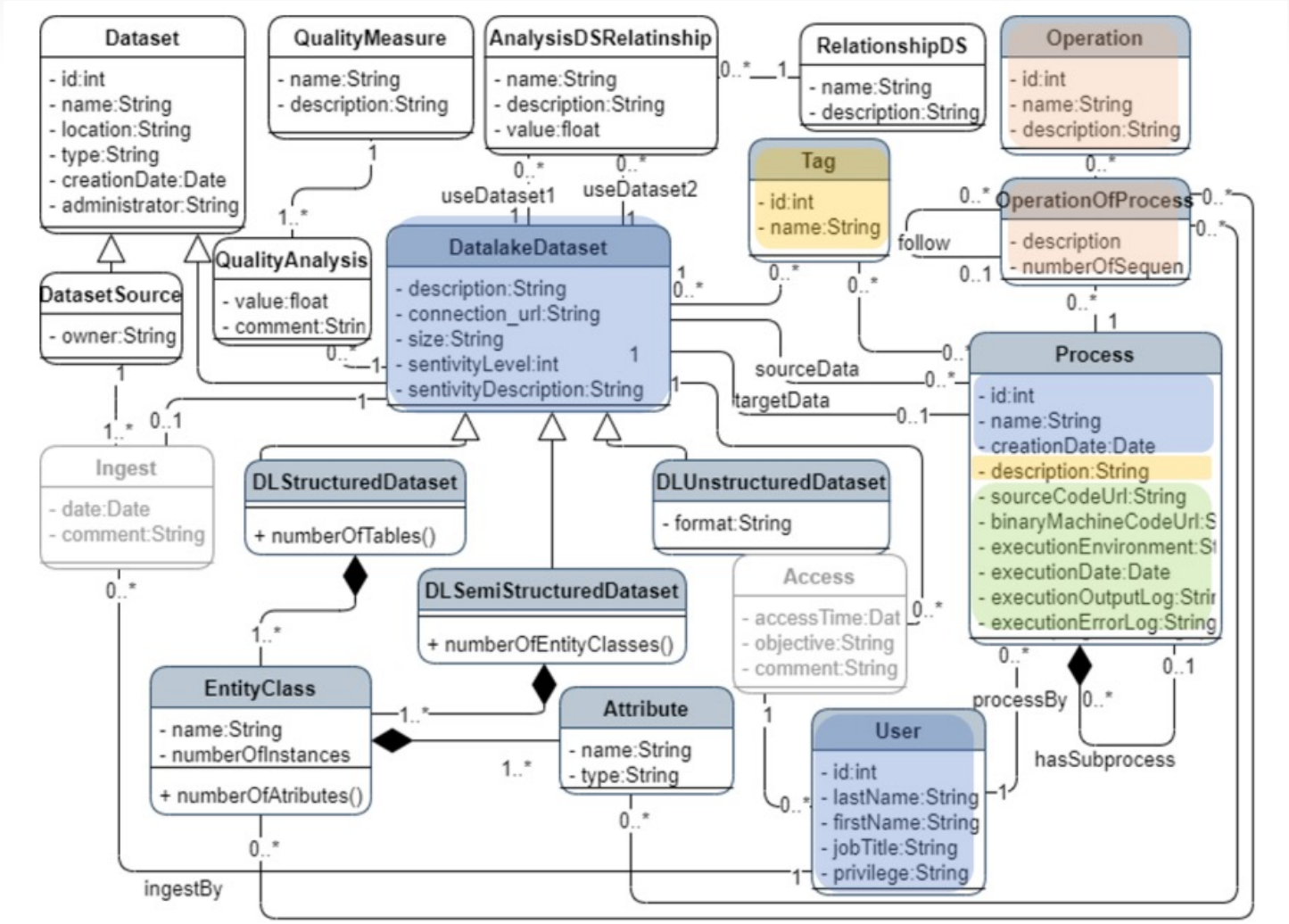
GEMMS

F. Ravat et al.

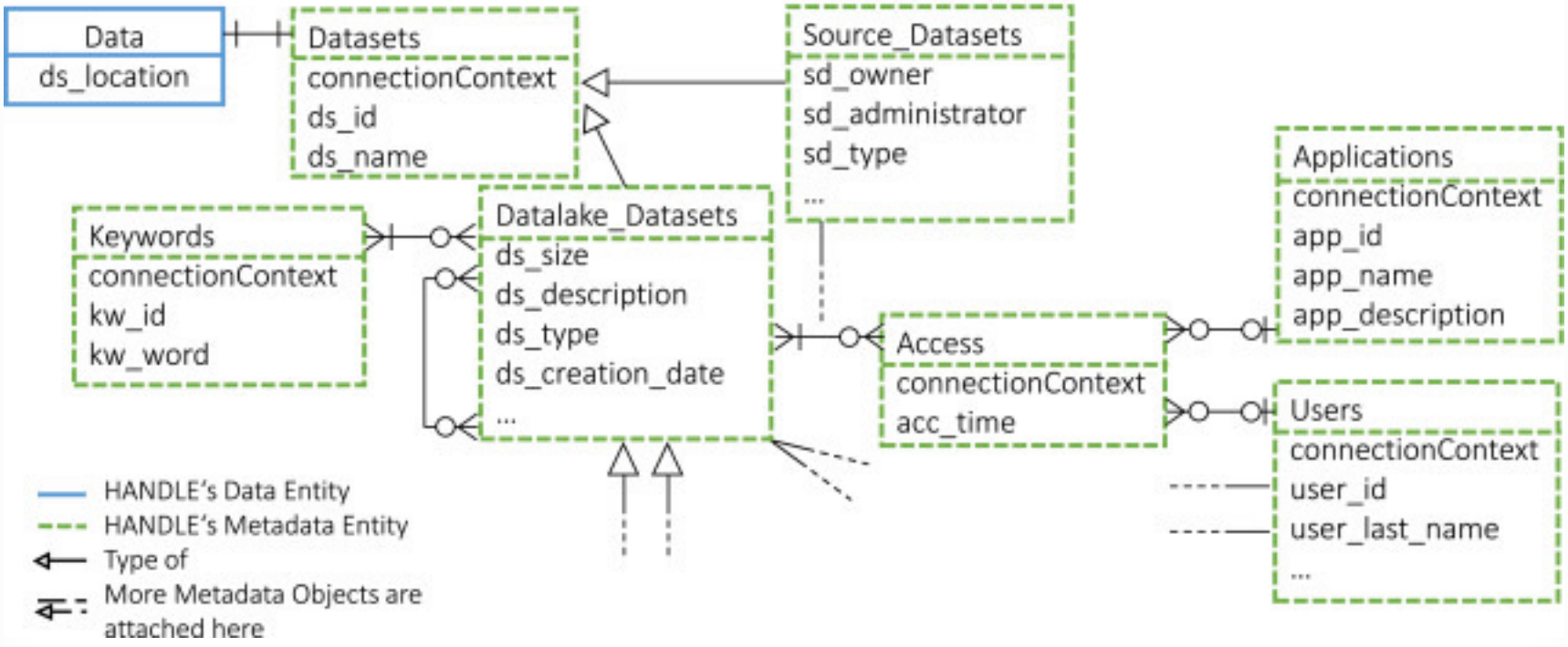
intra versus inter metadata: caracteristics, navigation, …

semantic (ontologies, taxonomies,…)

MEDAL and goldMEDAL –


Dataset
- id:int
- name:String
- location:String
- type:String
- creationDate:Date
- administrator:String
QualityMeasure
- name:String
- description:String
AnalysisDSRelatinship
- name:String
- description:String
- value:float
RelationshipDS
- name:String
- description:String
Operation
- id:int
- name:String
- description:String
Tag
- id:int
- name:String
OperationOfProcess
- description
- numberOfSequen
follow
useDataset1
useDataset2
QualityAnalysis
- value:float
- comment:Strin
DatalakeDataset
- description:String
- connection_url:String
- size:String
- sentivityLevel:int
- sentivityDescription:String
DatasetSource
- owner:String
sourceData
targetData
Process
- id:int
- name:String
- creationDate:Date
- description:String
- sourceCodeUrl:String
- binaryMachineCodeUrl:S
- executionEnvironment:St
- executionDate:Date
- executionOutputLog:Strir
- executionErrorLog:String
Ingest
- date:Date
- comment:String
DLStructuredDataset
+ numberOfTables()
DLUnstructuredDataset
- format:String
DLSemiStructuredDataset
+ numberOfEntityClasses()
Access
- accessTime:Dat
- objective:String
- comment:String
EntityClass
- name:String
- numberOfInstances
+ numberOfAtributes()
Attribute
- name:String
- type:String
User
- id:int
- lastName:String
- firstName:String
- jobTitle:String
- privilege:String
processBy
hasSubprocess
ingestBy

HANDLE mapped onto model by Ravat and Zhao



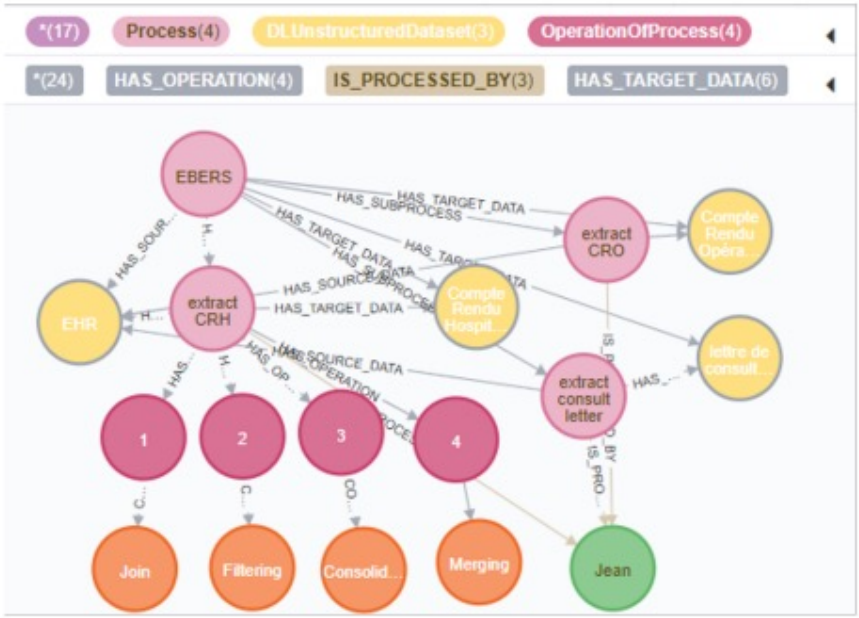
LINKING METADATA TO ONTOLOGIES AND KNOWLEDGE GRAPHS

We define a Semantic Data Lake as a tuple $SDL = \langle \mathcal{S}, \mathcal{G}, \mathcal{K}, m \rangle$, where $\mathcal{S} = \{S_1, \ldots, S_n\}$ is a set of data sources, $\mathcal{G} = \{G_1, \ldots, G_n\}$ is the corresponding set of metadata, $\mathcal{K}$ is a Knowledge Graph and $m \subseteq \mathcal{G} \times \mathcal{K}$ is a mapping function relating metadata to knowledge concepts. Our approach is agnostic w.r.t. both the degree of structuredness of the sources, ranging from structured datasets to semi-structured (e.g., XML, JSON) documents, and the specific DL architecture

Diamantini et al.

IMPLEMENTING META-DATA

More and more use of graph databases (e.g., Neo4J)

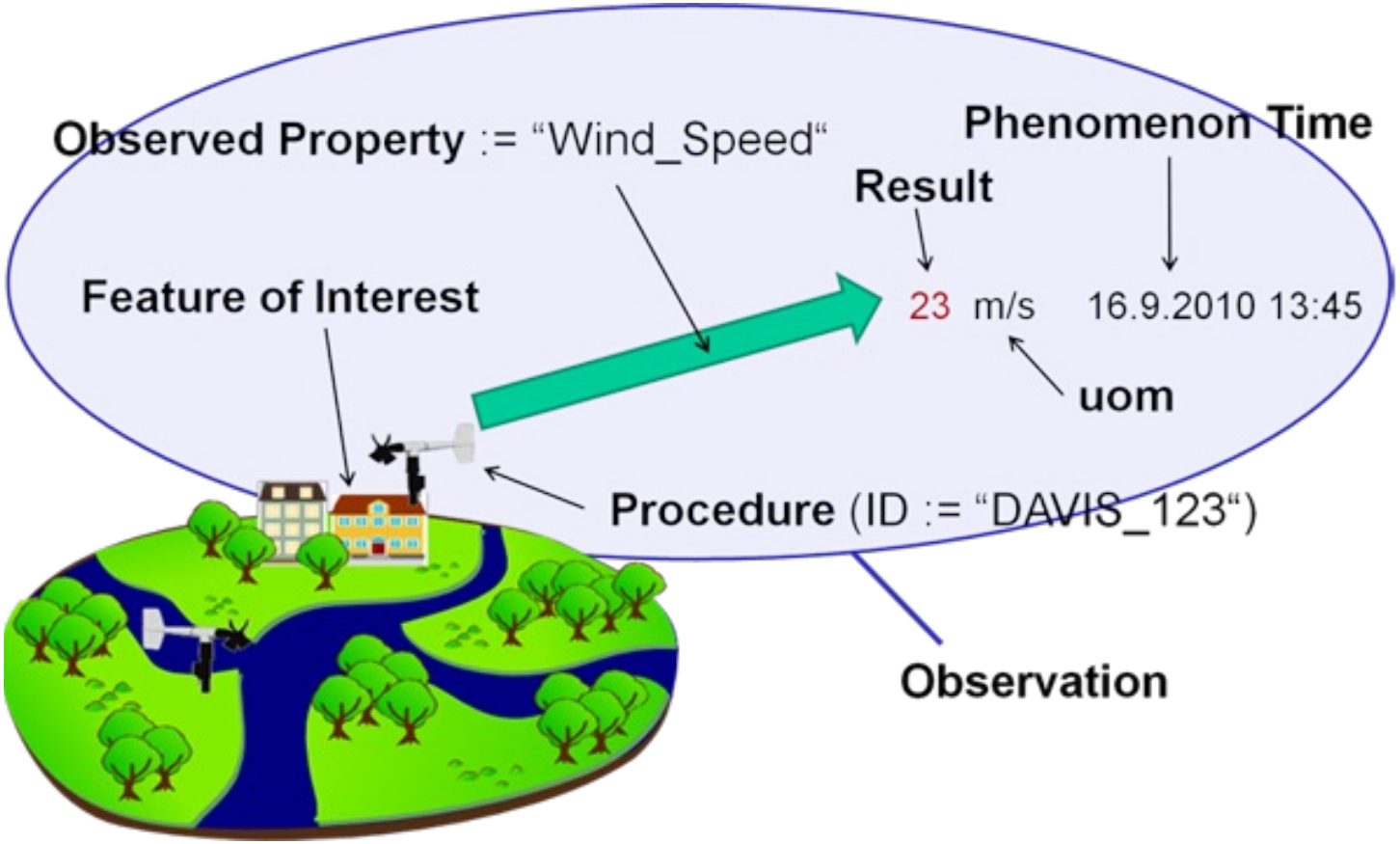



```
MATCH
    (p:Process)-[:HAS_TARGET_DATA]-(d),
    (p)-[:HAS_OPERATION]->(op:OperationOfProcess),
    (op)-[:CONCERNS_OPERATION]->(o:Operation),
    (p)-[:HAS_SUBPROCESS]-(p2:Process),
    (p)-[:IS_PROCESSED_BY]-(u),
    (p2)-[*1]-(n)
WHERE
    toLower(d.name) = 'compte rendu hospitalisation'
    OR toLower(d.name) = 'evaluation neuropsychologique'
    OR toLower(d.name) = 'examen neuropsychologique'
RETURN p,d,op,o,p2,u,n
```

Some Open Questions

- Hybridation of « data- » and « information-driven »
- Centralized versus decentralized architectures (data mesh, micro-services)
- Management of Constraints
- Data quality
- Easy navigation and exploitation
- Architectures and choices
  - files/DBMS
  - SQL/NoSQL
- Data Life Cycle, Storage/Backup/Archive
- …

Points of view and Raw Data



Death of Data

- Some contradictory injunctions?
  - e.g., Open Data versus environmental impact
- But also
  - some rules (e.g., GDPR)
  - Some constraints (data gravity)
- Should we keep all raw data even if it is cheap and « can be used in the future »?
- How to predict and measure data « usefulness » and reusability?
- …

REFERENCES

- F. Castanedo and S. Gidley, Understanding Metadata: Create the foundation for a Scalable Data Architecture, O'reilly, 2017.
- Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., and Martin, A. (2019). Life and death of data in data lakes: Preserving data usability and responsible governance. In Internet Science, Lecture Notes
- C. Diamantini, D. Potena, and E. Storti. A Knowledge-based approach to support analytic query answering in Semantic Data Lakes. ADBIS 2022.
- Dixon, J. (2010). Pentaho, Hadoop, and Data Lakes. Retrieved August 23, 2021, from https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and data-lakes/
- Inmon, B. (2016). Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump (1st ed.). Denville, NJ, USA: Technics Publications, LLC.
- Quix, C., Hai, R., Vatov, I. (2016). GEMMS: A Generic and Extensible Metadata Management System for data lakes. In CEUR Workshop Proceedings (Vol. 1612).
- Ravat, F., Zhao, Y. (2019b). Metadata Management for Data Lakes. In Communications in Computer and Information Science.
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).