



HAL
open science

FLOC : Un Système de Mesure Énergétique pour les infrastructures d'analyse de Big Data

Valera Humberto, Philippe Roose, Nathalie Vallès-Parlangeau

► To cite this version:

Valera Humberto, Philippe Roose, Nathalie Vallès-Parlangeau. FLOC : Un Système de Mesure Énergétique pour les infrastructures d'analyse de Big Data. Green Days 2024 "Explorer les multiples facettes de la sobriété numérique", Mar 2024, Toulouse, France. hal-04634200

HAL Id: hal-04634200

<https://univ-pau.hal.science/hal-04634200v1>

Submitted on 3 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FLOC: Un Système de Mesure Énergétique pour les infrastructures d'analyse de Big Data

Humberto VALERA (Domolandes),
Philippe ROOSE (LIUPPA),
Frank RAVAT (IRIT)
Jiefu SONG (IRIT)
Nathalie VALLES-PARLANGÉAU (LIUPPA)



Contexte

- Chaire Industrielle « **Bien vivre et bien vieillir** »
 - LIUPPA : 3 E/C
 - IRIT : 2 E/C
 - Domolandes : 1 Postdoc + 1 à venir
 - 3 PhD (dont 2 CIFRE)
 - Gestion et analyse écoresponsable de flux de données (CIFRE)
 - DataLake (plus) éco-responsable (CIFRE)
 - Trajectoires sémantiques du bien vivre et bien vieillir
 - + CC MACS; Région NA; + Entreprises (FMS, Digital Max, etc.)
 - Début 09/2022 (5 ans)

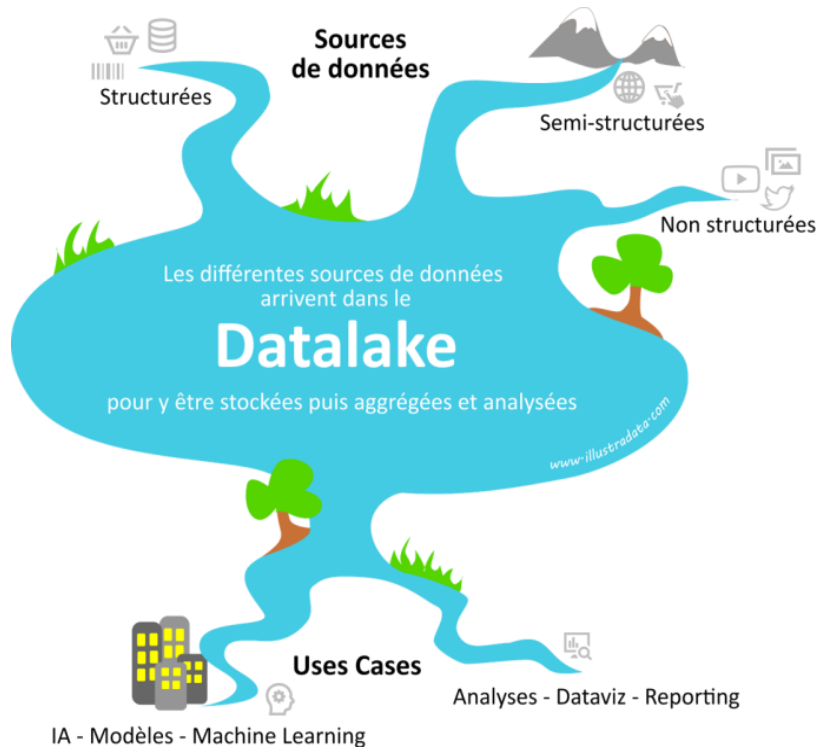


Constat

- Pour faire du « Bien vivre et bien vieillir », il faut
 - des données, beaucoup de données
 - **Collecter/Ingérer, Stocker, Analyser**
 - des Outils
 - Entrepôts de données – **Data Lakes**
 - **Pas très éco-responsable...** c'est même un peu orthogonal !
 - Mais alors...**Combien ça coute ?**



C'est quoi un lac de données?



Collecter/Ingérer; Stocker (dans format natif), Traiter **de grandes quantités de données**

⇒ Propose aussi:

- ⇒ Catalogue de métadonnées (qualité des données)
- ⇒ Politique et outils de gouvernance des données
- ⇒ Ouverture à tous types d'utilisateurs
- ⇒ Intégration de tous types de données
- ⇒ Organisation conceptuelle, logique et physique

Principaux objectifs d'amélioration des lacs de données aujourd'hui

Gestion des données, des métadonnées, de la qualité des données **sont prioritaires.**

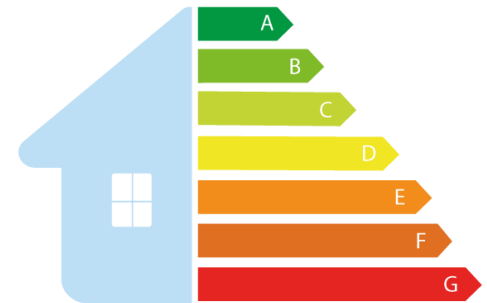
- Logique ELT (et pas ETL)

L'analyse de la consommation d'énergie de toutes ces opérations est souvent négligée.

- Négligée ? Non... Même pas abordée !

Et que dire de l'implémentation + déploiements

⇒ Efforts fait sur l'efficacité ... oui, mais pas énergétique



Que faut-il ?

- ⇒ **Collecte** : NET, RAM, HD
- ⇒ **Stockage** : HD
- ⇒ **Analyse** : CPU, RAM, HD



Enfin...pas tout à fait quand même !

- ⇒ Principaux outils existants tiennent compte du CPU, parfois un peu de la RAM (cache)... mais c'est tout.
- ⇒ **Il faut tenir compte du stockage (HD), du réseau (NET) et de l'usage de la RAM**

Bref...**Il nous faut du FLOC* !**

* Du pif...mais pas que !

Genèse...

- Thèse de doctorat d'Humberto Valera
- **PISCO** - An energy saving perspective for distributed environments: Deployment, scheduling and simulation with multidimensional entities for Software and Hardware
- Prototype:
 - Simulateur d'infra (Datacenter, PC, Smartphone, etc.) + Réseau
 - Déploiements dynamiques de MS en vue de diminuer la conso énergétique globale
- Breveté + Plusieurs transferts de techno (universités, entreprises)

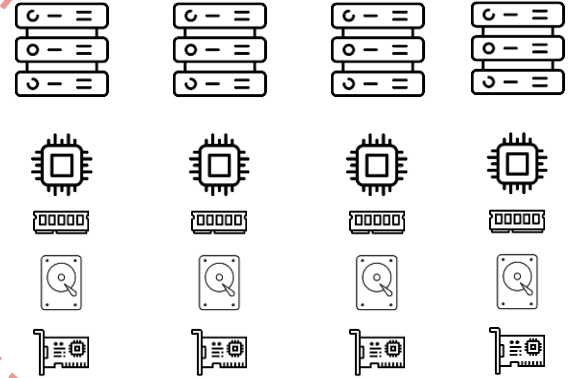
Donc, c'est quoi le premier pas vers un data lake? Mesurer avec FLOC!

FLOC évalue l'énergie des opérations sur chaque composants clés de chaque serveur.

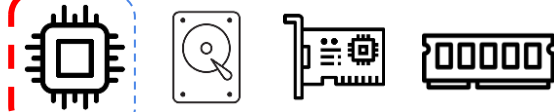
Obtention Σ énergie des opérations



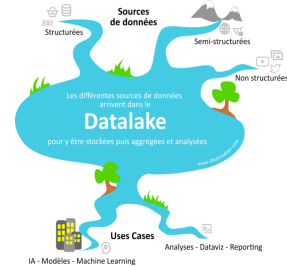
FLOC Analyse de tous les proc. et sous processus des opérations (ingestion, stockage, analyse).



Autres Outils

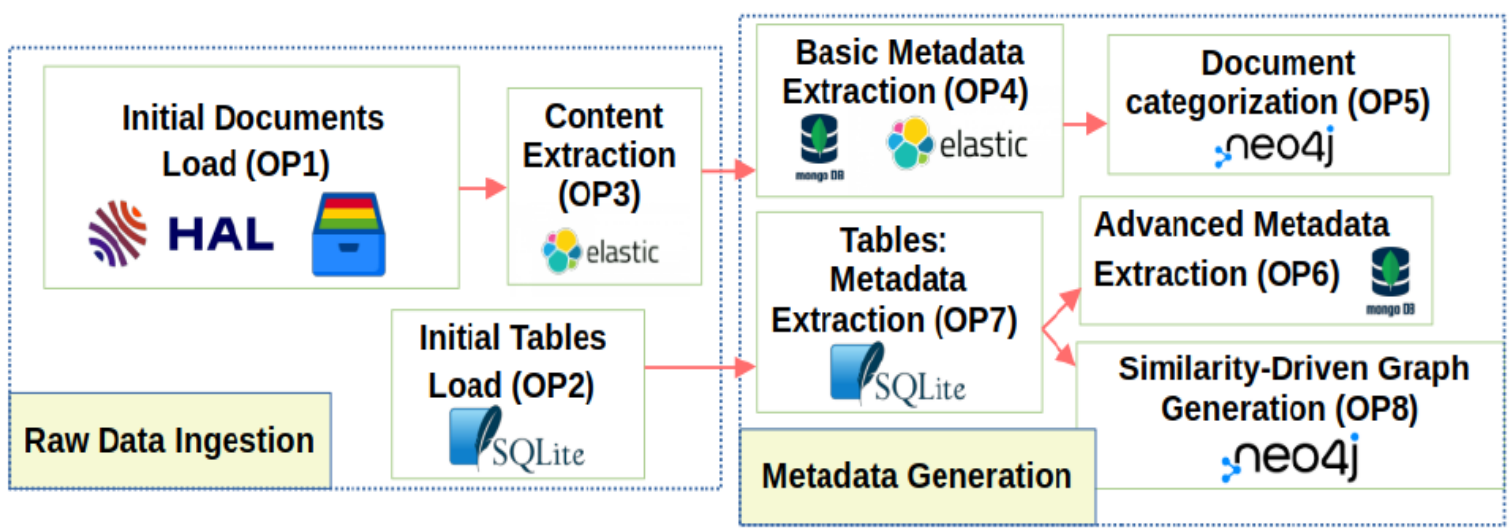


FLOC



Comment on a testé FLOC?





- Lac de données (**AUDAL***)
- Benchmark (**DLBench+**)
- ... depuis **min.io** est sorti (Infra DataLake + Benchmark)
- **Objectif**: mesurer l'efficacité des opérations (Opx) sur les données et les métadonnées.



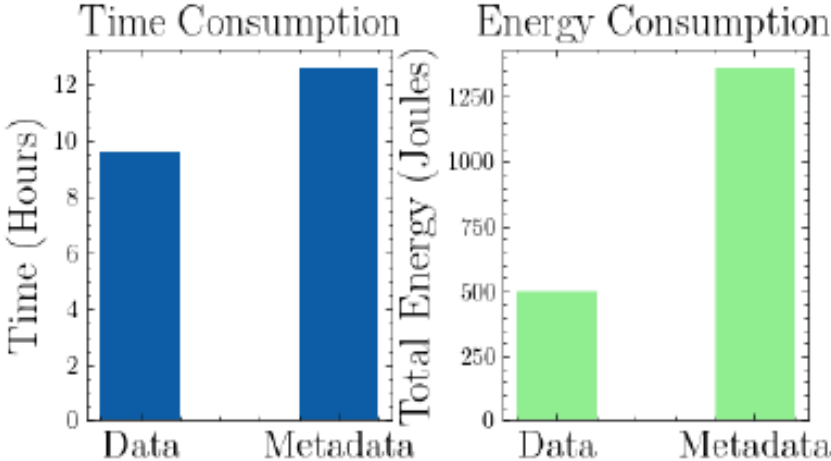
* P.N. Sawadogo, J. Darmont, C. Nous, « *Joint Management and Analysis of Textual Documents and Tabular Data within the AUDAL Data Lake* », 25th European Conference on Advances in Databases and Information Systems (ADBIS 2021)

Comment on a testé FLOC?

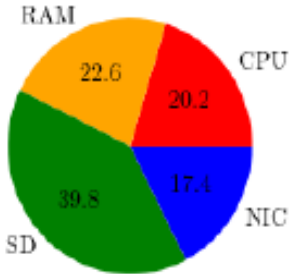
Etape 1 : Analyse de la consommation énergétique de la partie ingestion de DLBench+ dans un serveur avec les configurations suivantes:

HARDWARE SETUP		
Hardware	Spec.	Power Params.
CPU	Intel Core I7-850H 2.20GHZ (12 vcores)	TDP: 45W
RAM	samsung M471A2K43CB1-CRC	Values in section XX
NIC	Cannon Lake PCH CNVi WiFi	Download Power: 0.55W Upload Power: 1.029 W
SD	Samsung MZVLW512HMJP	Write Power: 6.1 W Read Power: 5.1 W
SOFTWARE SETUP		
Software	Spec.	
O.S.	Ubuntu Linux - Kernel V. 6.2.0-26	
Frameworks	  elasticsearch  mongoDB.  SQLite	

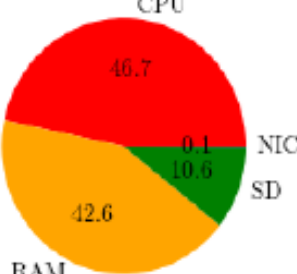
La consommation d'énergie: Data - Metadata



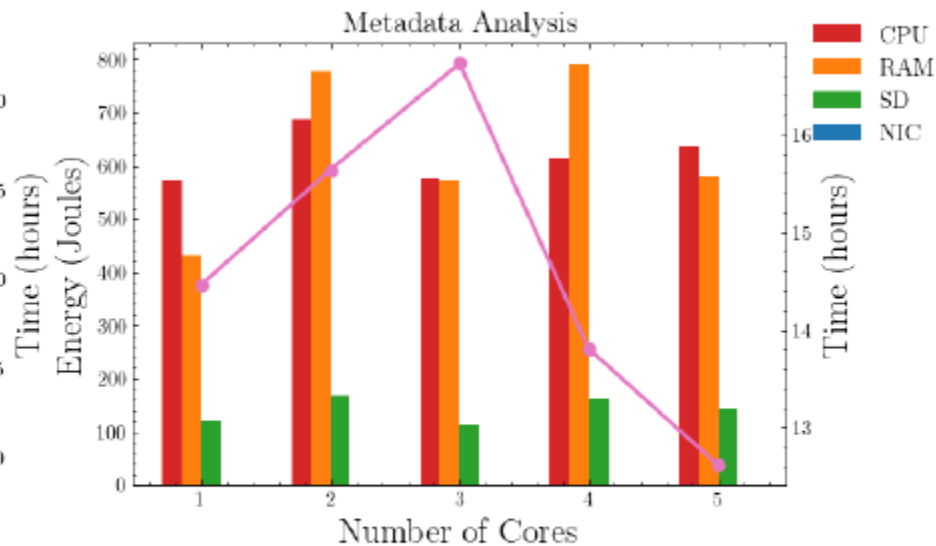
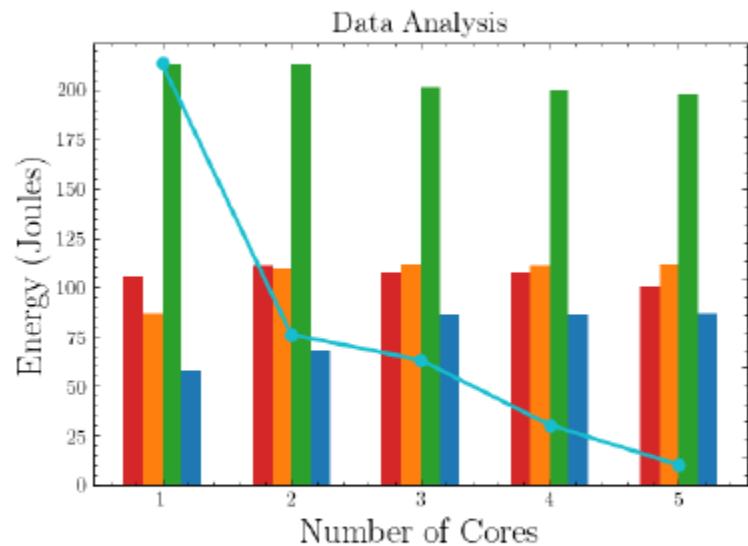
Data Processing



Metadata Processing



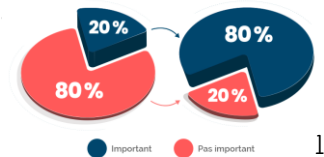
L'analyse du parallélisme





FLOC

- FLOC est générique
 - Créé pour...mais...pas que pour les Datalakes !
 - Open-source : <https://github.com/labDomolandes/FLOC>
 - Linux: *Arch, Ubuntu* (pour l'instant)
 - Mesure : **CPU, RAM, NET, HD (+GPU Nvidia à venir)**
 - Par appli (+sous processus), par processus, par liste de processus
 - Sur une durée en temps ou une durée d'exécution
 - Fourni des résultats pour identifier quels composants consomment le plus d'énergie au fil du temps
 - Focaliser sur les points de consommations => appliquer la **loi de Pareto**



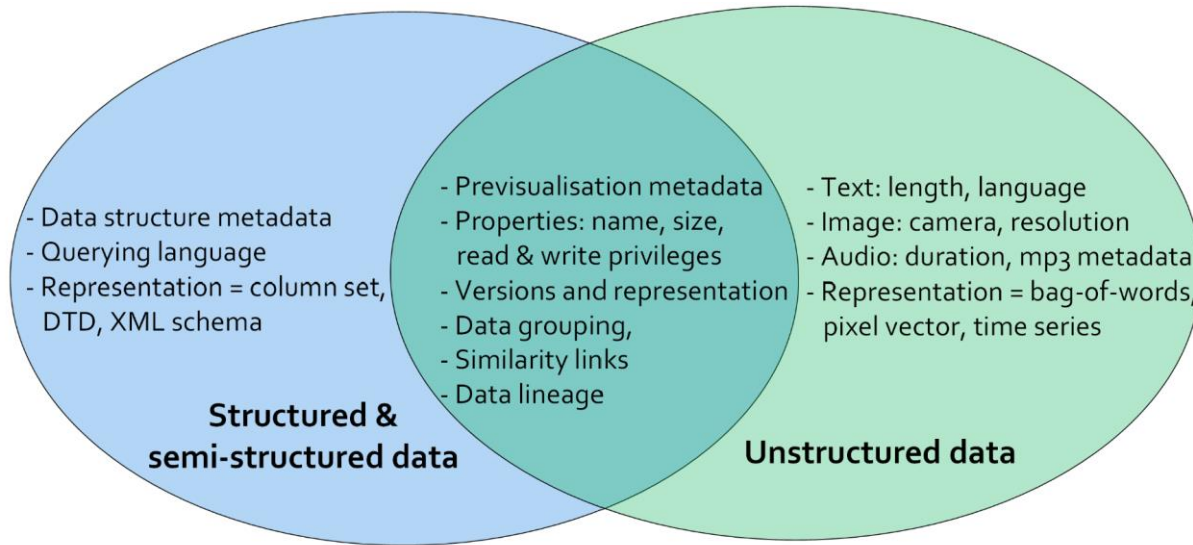
Meta-Data

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...

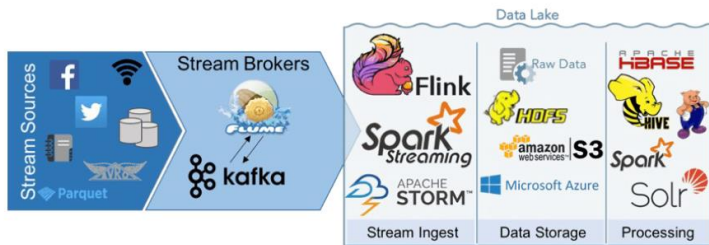
Métadonnées inter-objets	Exemples
Regroupements	Thématiques, par langue...
Similarités	Via des mesures de similarité
Parentés	Jointures, unions...

Métadonnées globales	Exemples
Ressources sémantiques	Ontologies, taxonomies...
Index	Index inversés
Journaux	Logs

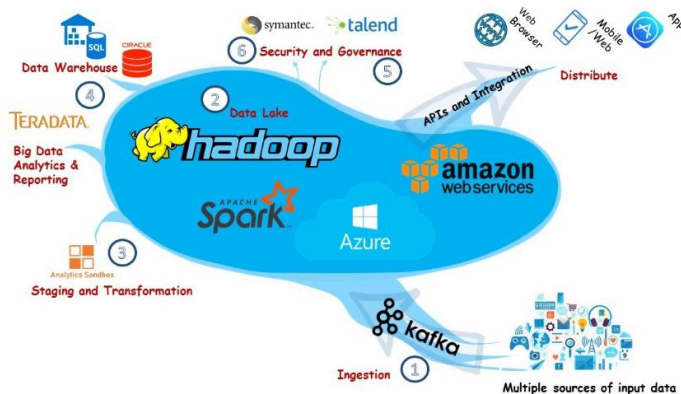
Structure des données



Techno



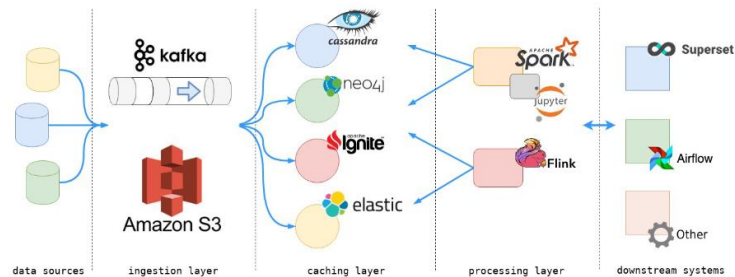
7wdata.be



kms-world.com

MC Common data lake technologies		Common formats
Metadata		
Storage		
Compute		Commonly managed by
(Logs)	Same as Storage 	

www.montecarlo.com



smartcat.io