



FONDATION
DE L'AP-HP

CHAIRE
AI-RACLÈS



PASC'22 Conference

Enabling Industrialized Analysis of Textual Documents in Data Lakes

Pegdwendé Nicolas SAWADOGO

sawadogonicholas44@gmail.com

June 27, 2022

Outline

- 1 Introduction
- 2 DL and DWH
- 3 AUDAL implementation
- 4 Textual analyses in AUDAL
- 5 Conclusion

Why Data Lakes?

Welcome to the big data era

Tremendous growth of produced and available data

Big data opportunities



Why Data Lakes?

Welcome to the big data era

Tremendous growth of produced and available data

Big data opportunities



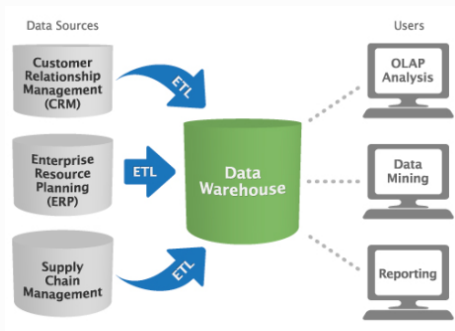
Big data challenges



Why Data Lakes?

From data warehouses to data lakes

- Data warehouses do achieve insights from big data.
- Distributed technologies to tackle **Volume**
- ...but **Variety** and **Velocity** pose great challenges.

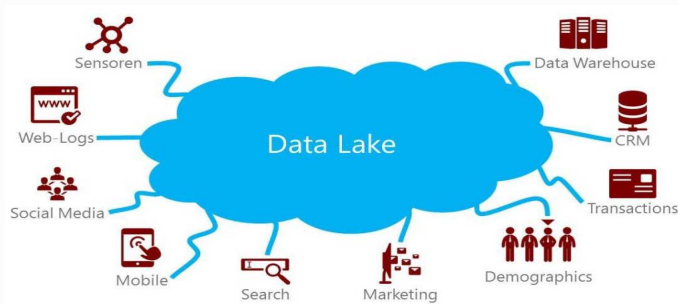


What is a Data Lake?

Definition

James Dixon (2010)

A data lake is a large **repository** of **raw** and **heterogeneous data**, fed by external sources and allowing users to **explore**, sample and **analyze** the data.

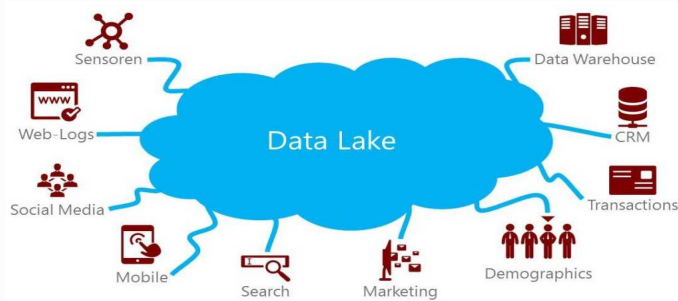


What is a Data Lake?

Definition

Sawadogo et al. (2019)

A data lake is a **scalable storage** and **analysis system** for **data** of any type, retained in their **native format** and used *mainly* by data **specialists** (statisticians, data scientists or analysts) for knowledge extraction.



Data Lake Issues

Avoiding the data swamp

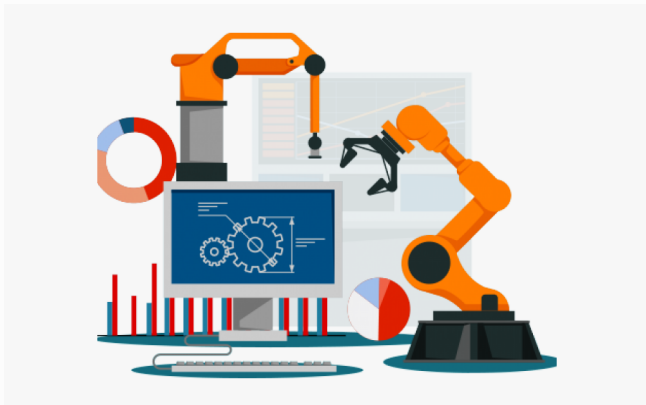
- **Schema-on-read** approach
- Efficient metadata system *required* for data access and querying
- ...but how to design such a system?



Data Lake Issues

Enabling industrialized analyses

- Open the data lake to business users
- Make easier advanced analyses
- Automate metadata management

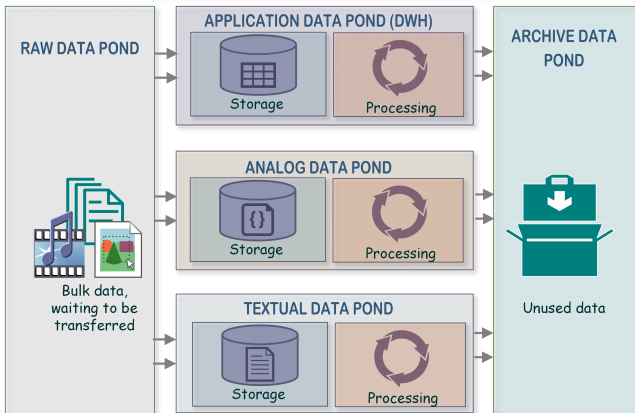


Outline

- 1 Introduction
- 2 DL and DWH**
- 3 AUDAL implementation
- 4 Textual analyses in AUDAL
- 5 Conclusion

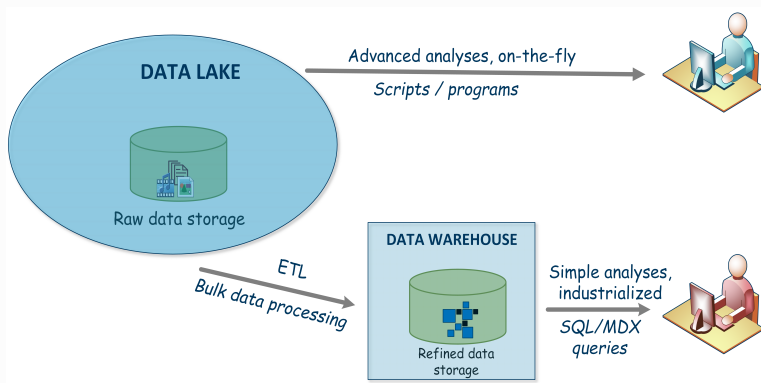
DWH in a DL

- Approach proposed by Inmon
- Induces a data siloing



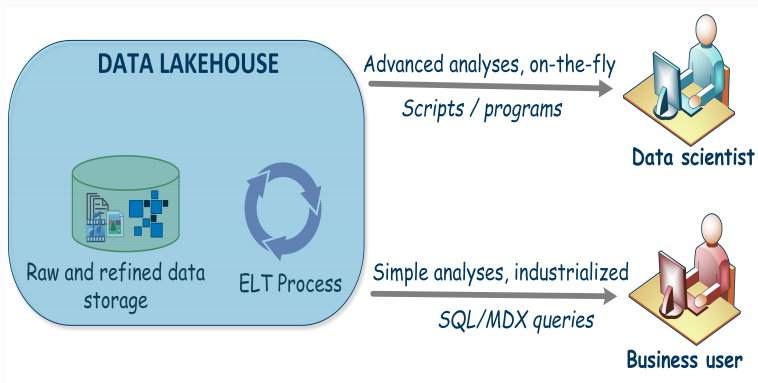
DL ahead of a DWH

- Based on a functional distinction
- DL for ponctual analyses, DWH for industrialized analyses



DL merged with a DWH

- Most recent approach (still maturing)
- Industrialization of analyses in the DL



Outline

- 1 Introduction
- 2 DL and DWH
- 3 AUDAL implementation**
- 4 Textual analyses in AUDAL
- 5 Conclusion

AURA-PMI Project

AURA-PMI, a multi-disciplinary project

- Research project in management
- Analysis of the digitization of small enterprises
- Comparison of digitization policies across categories of small enterprises



La Région
Auvergne-Rhône-Alpes

Coactis

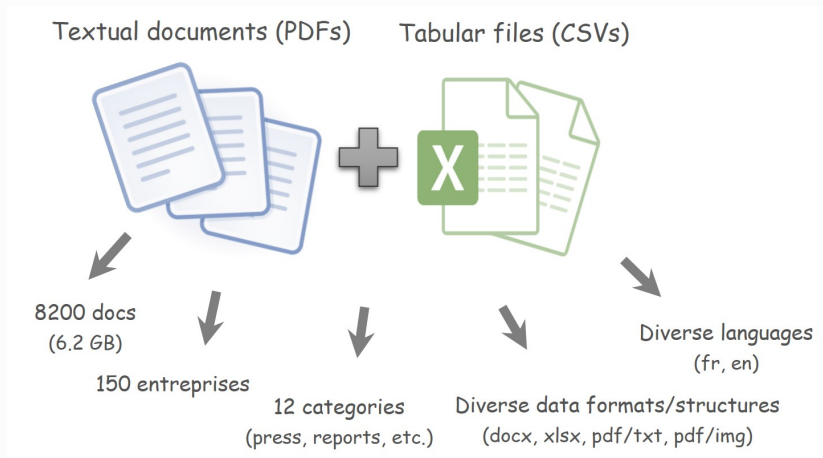


eric

AURA-PMI Project

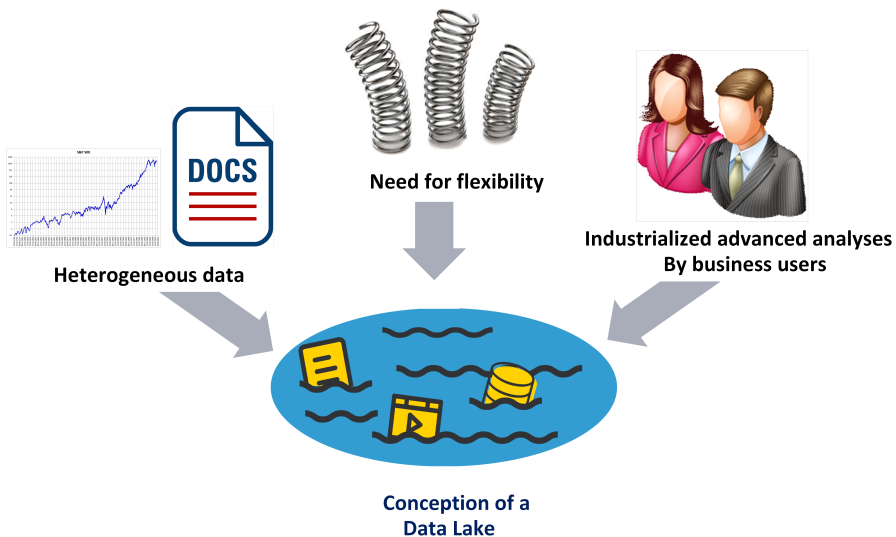
Extract insights from data

- Enterprises' characteristics (region, nb. employees, domain, etc.)
- Annual reports, financial reports, etc.

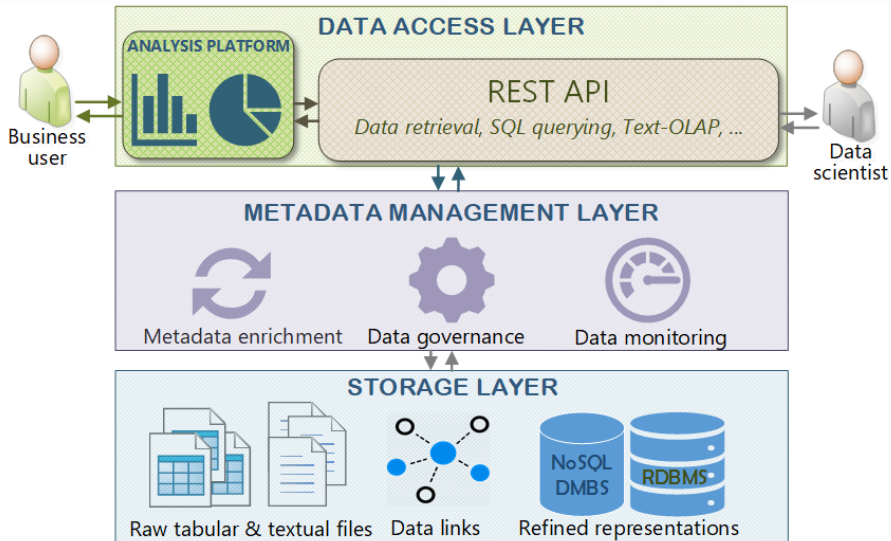


AURA-PMI Project

Need for a data lake



Architecture of AUDAL



Metadata management

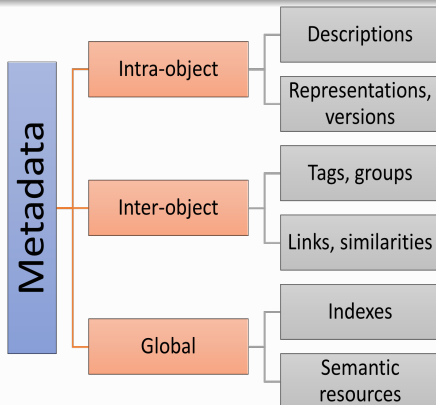
An extensive vision of metadata

Visengeriyeva (2020)

Metadata are **structured information** that **describes, explains, locates**, or otherwise makes it easier to **retrieve, use, or manage** information resources.

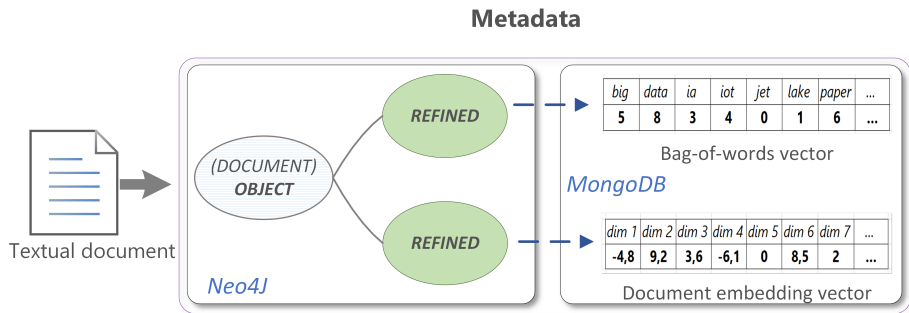
$$DL = \langle \mathcal{D}, \mathcal{M} \rangle$$

$$\mathcal{M} = \langle \mathcal{M}_{intra}, \mathcal{M}_{inter}, \mathcal{M}_{glob} \rangle$$



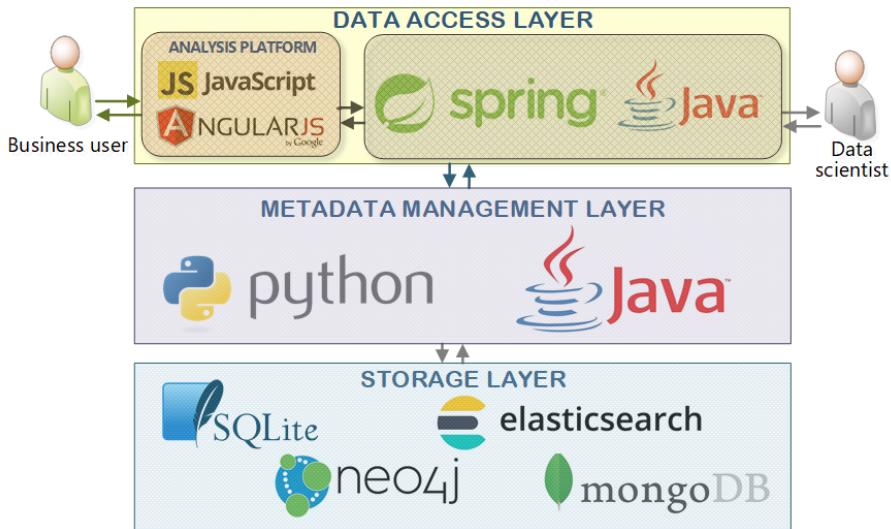
Metadata management in AUDAL (3/3)

Intra-object metadata



Technologies used in AUDAL

AUDAL = AUra-pmi DATA Lake



Outline

- 1 Introduction
- 2 DL and DWH
- 3 AUDAL implementation
- 4 Textual analyses in AUDAL**
- 5 Conclusion

Analyses dans AUDAL

AUDAL Analysis Interface

EXPLORATORY TASKS << DOCUMENT PROPERTIES (1) (2) (3) ANALYSES

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

#	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	👁

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- docCategorv

Analyses in AUDAL

Data retrieval

Keyword-based filtering

🔍 Terms filtering +

Terms

+ matching +

relation *client*

- matching +

Parameters

Strictness Any All

Fuzzy search Yes No

Terms extension None +

Category-based filtering

📁 Groupings -

Groups

- ▶ 1- cible
- ▶ 2- digitalNativity
- ▶ 3- docCategory
- ▶ 4- enterprise
- ▶ 5- language
- ▶ 6- mimeType
- ▶ 7- month
- ▼ 8- region
 - ALL
 - Auvergne-Rhone-Alpes [1342]
 - Bourgogne-Franche-Comte [404]
 - Bretagne [45]

Analyses in AUDAL

Textual data aggregation

Highlighting documents' content with a wordcloud



Analyses dans AUDAL

Textual data aggregations

Highlighting documents' content with a Concordance

2 - Comptes sociaux semestriels_2015_ANEVIA.pdf[10]

La valorisation des technologies acquises est amortie sur 8 ans et celle

d'utilisation retenues sont les suivantes : Nature Durée retenues ns Te
Relations

d'entreprises, nécessaires pour la mise en œuvre de partenariats OSE(
relations

clients correspondent aux portefeuilles **client**, stables et pérennes (cf

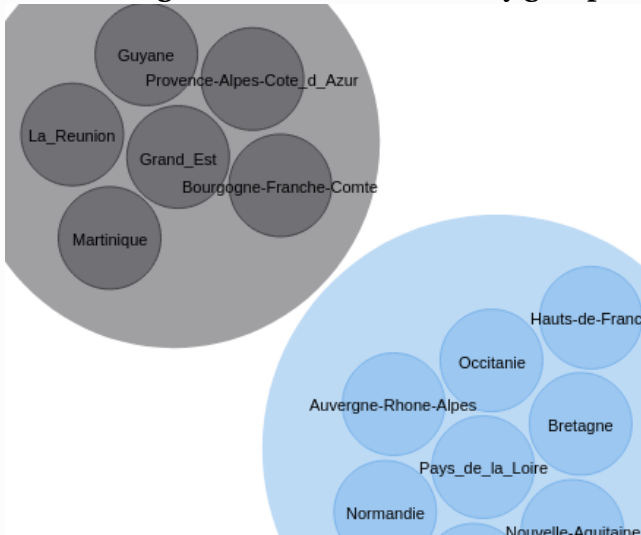
Relation Total acquises brevets clientèle Valeurs brutes a au 31.12.15

La juste valeur de la **relation** clientèle est évaluée selon la méthode di

Analyses in AUDAL

Textual data aggregation

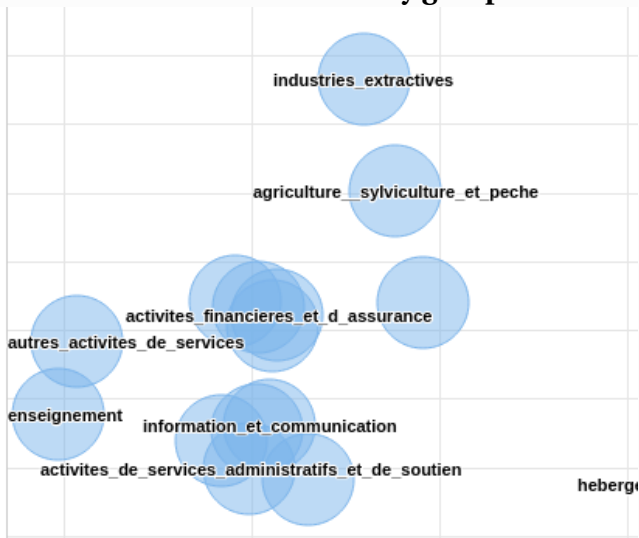
Clustering KMeans of documents by groups



Analyses in AUDAL

Textual data aggregation

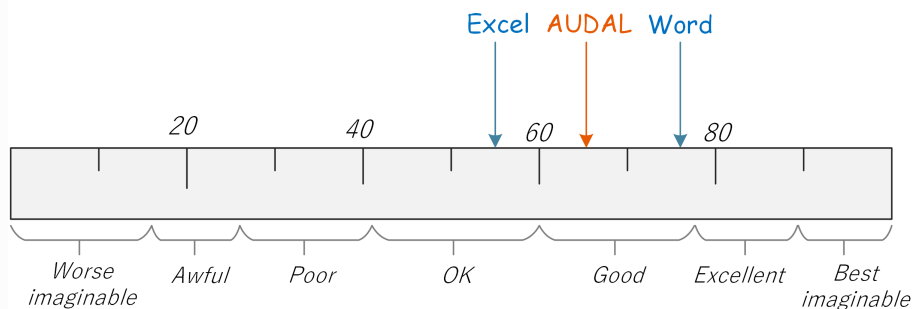
PCA of documents by groups



User experience with AUDAL

Usability

- ▶ Method *System Usability Scale* (SUS)
- ▶ Protocol based on a feedback from 6 users via a questionnaire



Outline

- 1 Introduction
- 2 DL and DWH
- 3 AUDAL implementation
- 4 Textual analyses in AUDAL
- 5 Conclusion**

Conclusion

We presented...

★ How DSS architectures evolved with the DL wave

- DWH in the DL
- DL ahead of the DWH
- DL merged with the DWH

★ How to activate industrialized analyses for textual documents in a DL

- Using an extensive vision of metadata
- Thanks to a principle of data polymorphism
- With a combination of storage technologies

Conclusion

We presented...

★ How DSS architectures evolved with the DL wave

- DWH in the DL
- DL ahead of the DWH
- DL merged with the DWH

★ How to activate industrialized analyses for textual documents in a DL

- Using an extensive vision of metadata
- Thanks to a principle of data polymorphism
- With a combination of storage technologies

What's next ?

▶ Activate deeper textual data analyses

- Sentiment analysis

▶ Industrialized analyses for more unstructured data in DLs

- Images
- Videos



FONDATION
DE L'AP-HP

CHAIRE
AI-RACLÈS



PASC'22 Conference

Enabling Industrialized Analysis of Textual Documents in Data Lakes

Pegdwendé Nicolas SAWADOGO

sawadogonicholas44@gmail.com

June 27, 2022