

Doctoral Consortium – ADBIS 2019 – *Bled, Slovenia*

Textual Data Analysis from Data Lakes

Pegdwendé N. Sawadogo

pegdwende.sawadogo@univ-lyon2.fr

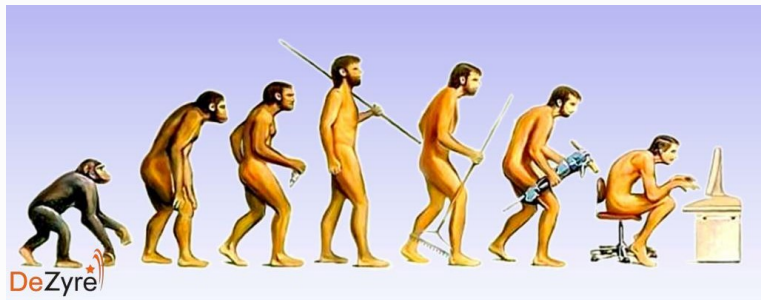
Supervised by Pr. Jérôme Darmont

September 8, 2019

Outline

- 1 Introduction
- 2 Thesis Objectives
- 3 Metadata Models
- 4 First Results
- 5 Conclusion

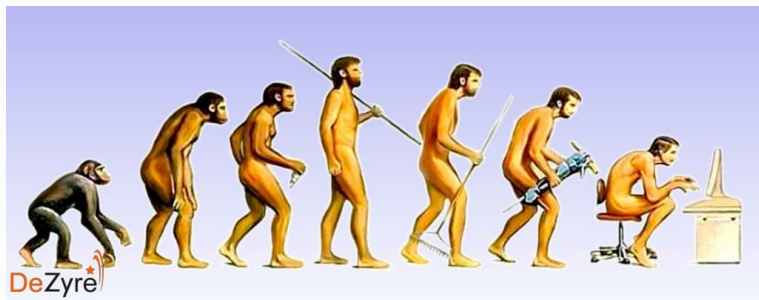
We are in big data era



innovations in IT until the 2000s

- / RDBMSs
- / World Wide Web
- / Data Warehouses

We are in big data era



innovations in IT until the 2000s

- / RDBMSs
- / World Wide Web
- / Data Warehouses

innovations in IT since the 2000s

- / NoSQL DBMSs
- / Internet of Things
- / Data Lakes

slideserve.com/DeZyre

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** for knowledge extraction.



Benefits of data lakes



Data governance



Dealing with unstructured data



Data polymorphism



Scalability



Cheap storage



Advanced analyses and KPIs

Data lakes challenges

"Data swamp" syndrome

- / Data swamp: inoperable DL
- / Poor metadata management
- / Poor data governance

medium.com



Data lakes challenges

"Data swamp" syndrome

- ✓ Data swamp: inoperable DL
- ✓ Poor metadata management
- ✓ Poor data governance

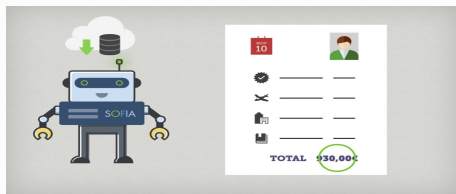
medium.com



Enabling industrialized analyses

- ✓ Opening DLs to business users
- ✓ Rich and intuitive metadata
- ✓ OLAP analysis

openflyers



- 1 Introduction
- 2 Thesis Objectives**
- 3 Metadata Models
- 4 First Results
- 5 Conclusion

Main Purposes

- ✓ Enable industrialized analyses from data lakes
- ✓ Focus on textual data analysis
- ✓ Alternative solution to text data warehouses

Main Purposes

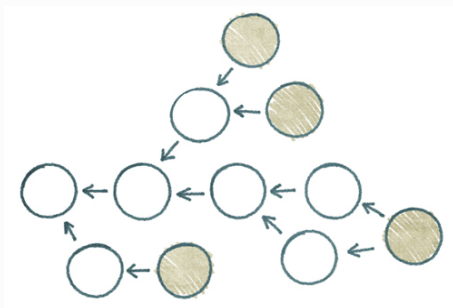
- ✓ Enable industrialized analyses from data lakes
- ✓ Focus on textual data analysis
- ✓ Alternative solution to text data warehouses



- 1 Introduction
- 2 Thesis Objectives
- 3 Metadata Models**
- 4 First Results
- 5 Conclusion

Data provenance-centric models

- ∩ DAG organization : nodes = data objects
- ∩ Vertices = operations (users, transformations, etc.)
- ∩ Help to understand, explain and repair inconsistencies in the data.

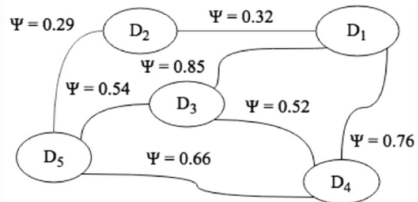


Similarity-centric models

- ✓ Allow to recommend related data
- ✓ Make it possible to detect data clusters

Simple variant

- Unoriented graph
- Nodes = data objects
- Edges = similarity strengths



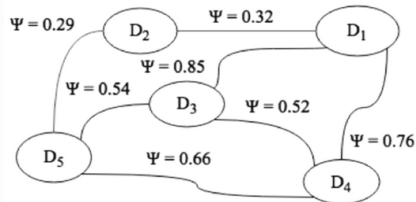
[Maccioni and Torlone, 2018]

Similarity-centric models

- ✓ Allow to recommend related data
- ✓ Make it possible to detect data clusters

Simple variant

- Unoriented graph
- Nodes = data objects
- Edges = similarity strengths



[Maccioni and Torlone, 2018]

Decomposition into droplets

- Data object = several nodes
- Connections are deduced from similarity between related “droplets”

Discussion (Sawadogo et al., 2019)

Metadata model/system	SE	DI	LG	DP	DV	UT
SPAR (Fauduet and Peyrard, 2010)	×	×	×			×
Terrizzano et al. (2015)	×	×			×	×
Singh et al. (2016)	×	×	×	×		
GOODS (Halevy et al., 2016)	×	×	×		×	×
Ground (Hellerstein et al., 2017)	×	×			×	×
KAYAK (Maccioni and Torlone, 2018)	×	×	×			
CoreKG (Beheshti et al., 2018)	×	×	×	×		×
Diamantini et al. (2018)	×		×	×		

*SE: Semantic Enrichment - DI: Data Indexing - LG: Links Generation
DP: Data Polymorphism - DV: Data Versioning - UT: Usage Tracking*

[Sawadogo et al., 2019b] - BBIGAP@ADBIS 2019

Discussion (Sawadogo et al., 2019)

Metadata model/system	SE	DI	LG	DP	DV	UT
SPAR (Fauduet and Peyrard, 2010)	×	×	×			×
Terrizzano et al. (2015)	×	×			×	×
Singh et al. (2016)	×	×	×	×		
GOODS (Halevy et al., 2016)	×	×	×		×	×
Ground (Hellerstein et al., 2017)	×	×			×	×
KAYAK (Maccioni and Torlone, 2018)	×	×	×			
CoreKG (Beheshti et al., 2018)	×	×	×	×		×
Diamantini et al. (2018)	×		×	×		

SE: Semantic Enrichment - DI: Data Indexing - LG: Links Generation
DP: Data Polymorphism - DV: Data Versioning - UT: Usage Tracking

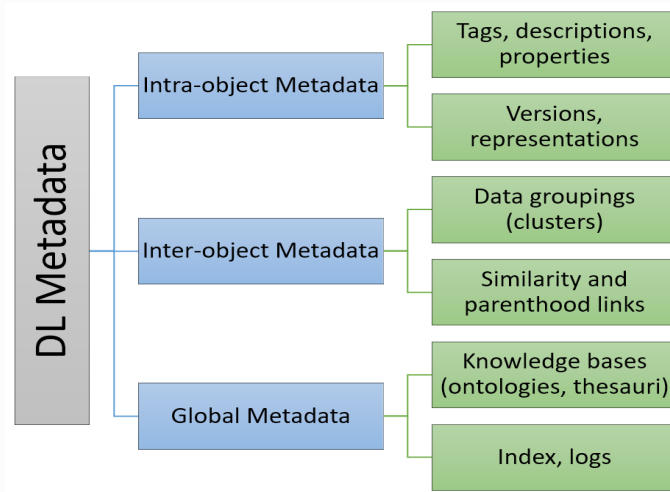
[Sawadogo et al., 2019b] - BBIGAP@ADBIS 2019

∩ No comprehensive metadata model

∩ Data versioning and data polymorphism as advanced features

- 1 Introduction
- 2 Thesis Objectives
- 3 Metadata Models
- 4 First Results**
- 5 Conclusion

Typology of data lake metadata



[Sawadogo et al., 2019a] - ICEIS 2019

Generic metadata model for data lakes

Intra-objects metadata

Generic metadata model for data lakes

Intra-objects metadata

Inter-objects metadata

Generic metadata model for data lakes

Intra-objects metadata

Inter-objects metadata

Generic metadata model for data lakes

Intra-objects metadata

Inter-objects metadata

Global metadata

- ∕ Not included
- ∕ Ontologies = graphs
- ∕ Mostly depend on adopted technologies

Expected features

/ Data search

- keyword/pattern-based querying
- Query extension
- Navigation accross data

Expected features

/ Data search

- keyword/patern-based querying
- Query extension
- Navigation accross data

/ Navigation/OLAP analysis

- Dimensions = data groupings
- Hierarchies = ontologies
- Aggregations = data fusion

Expected features

/ Data search

- keyword/patern-based querying
- Query extension
- Navigation accross data

/ Navigation/OLAP analysis

- Dimensions = data groupings
- Hierarchies = ontologies
- Aggregations = data fusion

/ Recommendation of data

- Similar data
- Affiliated data
- Data of same cluster

Expected features

/ Data search

- keyword/patern-based querying
- Query extension
- Navigation accross data

/ Navigation/OLAP analysis

- Dimensions = data groupings
- Hierarchies = ontologies
- Aggregations = data fusion

/ Recommendation of data

- Similar data
- Affiliated data
- Data of same cluster

/ Compliant with FAIR principles

- Findable
- Accessible
- Interoperable
- Re-usable

- 1 Introduction
- 2 Thesis Objectives
- 3 Metadata Models
- 4 First Results
- 5 Conclusion**

Conclusion

- **Overview**

- ✓ Opening data lakes to business users
- ✓ 6 key features to evaluate data lakes metadata models/systems
- ✓ Consideration of OLAP analysis in data lakes

Conclusion

- **Overview**

- ✓ Opening data lakes to business users
- ✓ 6 key features to evaluate data lakes metadata models/systems
- ✓ Consideration of OLAP analysis in data lakes

- **Future works**

- ✓ Implementing our metadata model into a metadata system
- ✓ Designing an OLAP analysis platform for textual data ponds
- ✓ Identifying techniques and tools to ensure scalability

Doctoral Consortium – ADBIS 2019 – *Bled, Slovenia*

Textual Data Analysis from Data Lakes

Pegdwendé N. Sawadogo

pegdwende.sawadogo@univ-lyon2.fr

Supervised by Pr. Jérôme Darmont

September 8, 2019