Metadata systems for data lakes : models and features

BBIGAP Workshop @ ADBIS 2019

September 8th, 2019 in Bled, Slovenia

P. N. Sawadogo[1], E. Scholly[1,2], C. Favre[1], E. Ferey[2], S. Loudcher[1], J. Darmont[1]

[1]University of Lyon, Lyon 2, ERIC EA 3083
[2]BIAL-X

Introduction

Context

A disruption : Big Data

[Miloslavskaya and Tolstoy, 2016, Hai et al., 2016, Alrehamy and Walker, 2015, Terrizzano et al., 2015]

Context

A disruption : Big Data

Multiple issues : Volume, Variety, Velocity...

[Miloslavskaya and Tolstoy, 2016, Hai et al., 2016, Alrehamy and Walker, 2015, Terrizzano et al., 2015]

Context

A disruption : Big Data

Multiple issues : Volume, Variety, Velocity...

A solution (for Big Data management) : Data Lakes

[Miloslavskaya and Tolstoy, 2016, Hai et al., 2016, Alrehamy and Walker, 2015, Terrizzano et al., 2015]

Context

A disruption : Big Data

Multiple issues : Volume, Variety, Velocity...

A solution (for Big Data management) : Data Lakes
$\rightarrow$ integrated data storage without predefined schema

[Miloslavskaya and Tolstoy, 2016, Hai et al., 2016, Alrehamy and Walker, 2015, Terrizzano et al., 2015]

Context

A disruption : Big Data

Multiple issues : Volume, Variety, Velocity...

A solution (for Big Data management) : Data Lakes
$\rightarrow$ integrated data storage without predefined schema

Research problems : metadata systems for Data Lakes

[Miloslavskaya and Tolstoy, 2016, Hai et al., 2016, Alrehamy and Walker, 2015, Terrizzano et al., 2015]

Context

A disruption : Big Data

Multiple issues : Volume, Variety, Velocity...

A solution (for Big Data management) : Data Lakes
$\rightarrow$ integrated data storage without predefined schema

Research problems : metadata systems for Data Lakes

- Organization ?
- Efficiency evaluation ?

[Miloslavskaya and Tolstoy, 2016, Hai et al., 2016, Alrehamy and Walker, 2015, Terrizzano et al., 2015]

Table of contents

Data Lake Concept

What is a data lake ?

[Dixon, 2010]

A data lake is a large repository of heterogeneous raw data, supplied by external data sources and from which various analyses can be performed.

[Fang, 2015, O'Leary, 2014, Sirosh, 2016, Khine and Wang, 2017]

What is a data lake ?

[Dixon, 2010]

A data lake is a large repository of heterogeneous raw data, supplied by external data sources and from which various analyses can be performed.

An alternative to data marts / data warehouses

[Fang, 2015, O'Leary, 2014, Sirosh, 2016, Khine and Wang, 2017]

What is a data lake ?

[Dixon, 2010]

A data lake is a large repository of heterogeneous raw data, supplied by external data sources and from which various analyses can be performed.

An alternative to data marts / data warehouses

- *Schema-on-read*

[Fang, 2015, O'Leary, 2014, Sirosh, 2016, Khine and Wang, 2017]

What is a data lake ?

[Dixon, 2010]

A data lake is a large repository of heterogeneous raw data, supplied by external data sources and from which various analyses can be performed.

An alternative to data marts / data warehouses

- *Schema-on-read*
- Data variety

[Fang, 2015, O'Leary, 2014, Sirosh, 2016, Khine and Wang, 2017]

What is a data lake ?

[Dixon, 2010]

A data lake is a large repository of heterogeneous raw data, supplied by external data sources and from which various analyses can be performed.

An alternative to data marts / data warehouses

- *Schema-on-read*
- Data variety

[Fang, 2015, O'Leary, 2014, Sirosh, 2016, Khine and Wang, 2017]

What is a data lake ?

[Dixon, 2010]

A data lake is a large repository of heterogeneous raw data, supplied by external data sources and from which various analyses can be performed.

An alternative to data marts / data warehouses

- *Schema-on-read*
- Data variety

Not only Hadoop

[Fang, 2015, O'Leary, 2014, Sirosh, 2016, Khine and Wang, 2017]

A recent, more complete definition

[Madera and Laurent, 2016]

A data lake is a logical view of all data sources and datasets in their raw format, accessible by data scientists or statisticians for knowledge extraction.

Definition complemented by :

A recent, more complete definition

[Madera and Laurent, 2016]

A data lake is a logical view of all data sources and datasets in their raw format, accessible by data scientists or statisticians for knowledge extraction.

Definition complemented by :

1. a set of metadata ;
2. data governance policy tools ;
3. limited to statisticians and data scientists ;
4. integrates data of all types and formats ;
5. logical and physical organization.

Our definition

Definition

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.

Its characteristics include :

Our definition

Definition

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.

Its characteristics include :

1. a metadata catalog ;
2. data governance policies and tools ;
3. accessibility to various kinds of users ;
4. integration of any type of data ;
5. a logical and physical organization ;
6. scalability.

Features of a metadata system

Expected features

6 features :

Expected features

6 features :

- **Semantic enrichment** : description of the context of data

[Ansari et al., 2018, Singh et al., 2016, Maccioni and Torlone, 2018, Farrugia et al., 2016, Diamantini et al., 2018]

Expected features

6 features :

- **Semantic enrichment** : description of the context of data
- **Data indexing** : retrieve datasets based on specific characteristics

[Ansari et al., 2018, Singh et al., 2016, Maccioni and Torlone, 2018, Farrugia et al., 2016, Diamantini et al., 2018]

Expected features

6 features :

- **Semantic enrichment** : description of the context of data
- **Data indexing** : retrieve datasets based on specific characteristics
- **Link generation** : similarity relationships or preexisting links between datasets

[Ansari et al., 2018, Singh et al., 2016, Maccioni and Torlone, 2018, Farrugia et al., 2016, Diamantini et al., 2018]

Expected features

6 features :

- **Semantic enrichment** : description of the context of data
- **Data indexing** : retrieve datasets based on specific characteristics
- **Link generation** : similarity relationships or preexisting links between datasets
- **Data polymorphism** : multiple representations of the same data

[Ansari et al., 2018, Singh et al., 2016, Maccioni and Torlone, 2018, Farrugia et al., 2016, Diamantini et al., 2018]

Expected features

6 features :

- **Semantic enrichment** : description of the context of data
- **Data indexing** : retrieve datasets based on specific characteristics
- **Link generation** : similarity relationships or preexisting links between datasets
- **Data polymorphism** : multiple representations of the same data
- **Data versionning** : data changes while conserving previous states

[Ansari et al., 2018, Singh et al., 2016, Maccioni and Torlone, 2018, Farrugia et al., 2016, Diamantini et al., 2018]

Expected features

6 features :

- **Semantic enrichment** : description of the context of data
- **Data indexing** : retrieve datasets based on specific characteristics
- **Link generation** : similarity relationships or preexisting links between datasets
- **Data polymorphism** : multiple representations of the same data
- **Data versionning** : data changes while conserving previous states
- **Usage tracking** : interactions between users and the lake

[Ansari et al., 2018, Singh et al., 2016, Maccioni and Torlone, 2018, Farrugia et al., 2016, Diamantini et al., 2018]

Comparison of metadata systems

System	Type	SE	DI	LG	DP	DV	UT
SPAR (Fauduet and Peyrard, 2010)	♦♯	✓	✓	✓			✓
Alrehamy and Walker (2015)	♦	✓		✓			
Terrizzano et al. (2015)	♦	✓	✓			✓	✓
Constance (Hai et al., 2016)	♦	✓	✓				
GEMMS (Quix et al., 2016)	◊	✓					
CLAMS (Farid et al., 2016)	♦	✓					
Suriarachchi and Plale (2016)	◊				✓		✓
Singh et al. (2016)	♦	✓	✓	✓	✓		
Farrugia et al. (2016)	♦			✓			
GOODS (Halevy et al., 2016)	♦	✓	✓	✓		✓	✓
CoreDB (Beheshti et al., 2017)	♦		✓				✓
Ground (Hellerstein et al., 2017)	◊♯	✓	✓			✓	✓
KAYAK (Maccioni and Torlone, 2018)	♦	✓	✓	✓			
CoreKG (Beheshti et al., 2018)	♦	✓	✓	✓	✓		✓
Diamantini et al. (2018)	◊	✓		✓	✓		

♦ : Data lake implementation ◊ : Metadata model
♯ : Model or implementation assimilable to a data lake

[Hellerstein et al., 2017, Beheshti et al., 2018, Beheshti et al., 2017, Quix et al., 2016, Halevy et al., 2016, Farid et al., 2016, Suriarachchi and Plale, 2016]

Metadata typology and MEDAL

The object notion

Object = a set of homogeneous data

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

- Intra-object metadata : related to one object

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

- Intra-object metadata : related to one object
- Inter-object metadata : relations between objects

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

- Intra-object metadata : related to one object
- Inter-object metadata : relations between objects
- Global metadata : general context

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

- Intra-object metadata : related to one object
- Inter-object metadata : relations between objects
- Global metadata : general context

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

- Intra-object metadata : related to one object
- Inter-object metadata : relations between objects
- Global metadata : general context

MEtadata model for DAta Lakes (MEDAL) : a graph-based modeling

The object notion

Object = a set of homogeneous data

Multiple terms proposed : data units, entities, datasets...

Three categories of metadata :

- Intra-object metadata : related to one object
- Inter-object metadata : relations between objects
- Global metadata : general context

MEtadata model for DAta Lakes (MEDAL) : a graph-based modeling
$\rightarrow$ Object = **Hypernode**

Data Lake definition

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

- $\mathcal{D}$ a set of raw data
- $\mathcal{M}$ a set of metadata

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

- $\mathcal{M}_{intra}$ the set of intra-object metadata
- $\mathcal{M}_{inter}$ the set of inter-object metadata
- $\mathcal{M}_{glob}$ the set of global metadata

Intra-object metadata

- **Properties** : general description

Intra-object metadata

- **Properties** : general description
- **Summaries and previews** : overview of the content or structure

Intra-object metadata

- **Properties** : general description
- **Summaries and previews** : overview of the content or structure
- **Versions and representations** : updated and reformated data

Intra-object metadata

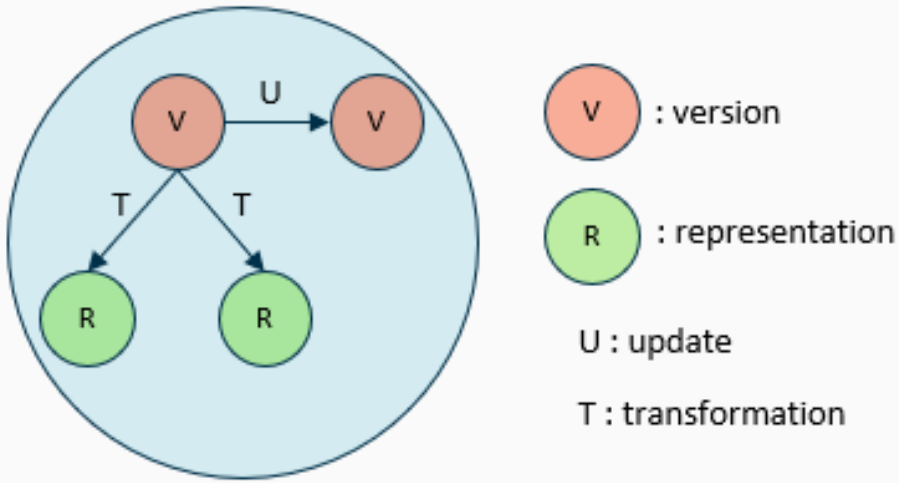
- **Properties** : general description
- **Summaries and previews** : overview of the content or structure
- **Versions and representations** : updated and reformated data
- **Semantic metadata** : annotations that help understand the meaning of data

Intra-object metadata

- **Properties** : general description
- **Summaries and previews** : overview of the content or structure
- **Versions and representations** : updated and reformated data
- **Semantic metadata** : annotations that help understand the meaning of data

Intra-object metadata

- **Properties** : general description
- **Summaries and previews** : overview of the content or structure
- **Versions and representations** : updated and reformated data
- **Semantic metadata** : annotations that help understand the meaning of data



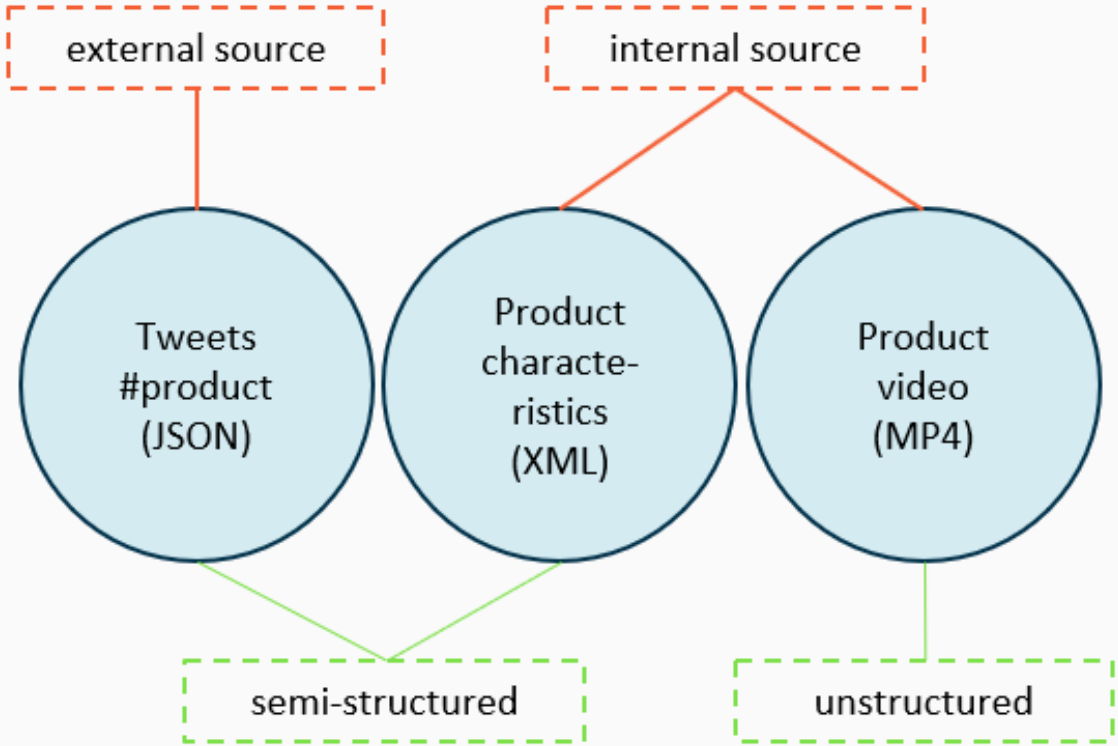
In MEDAL : nodes and directed edges (+ attributes)

Inter-object metadata

- **Objects groupings** : organize objects into collections

Inter-object metadata

- **Objects groupings** : organize objects into collections



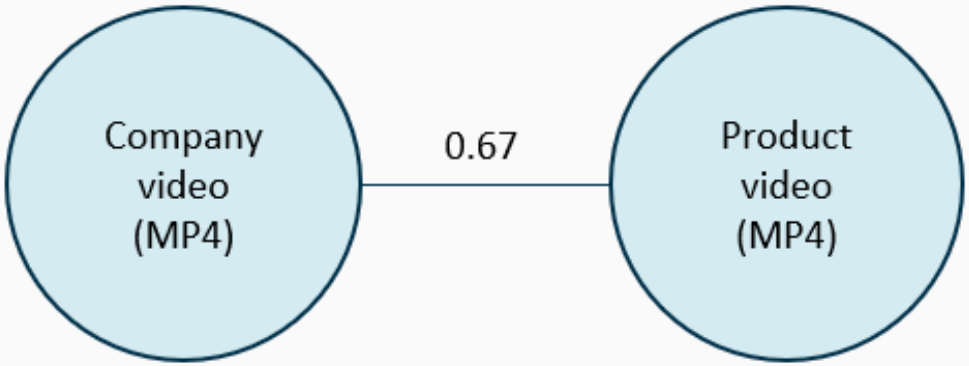
In MEDAL : hyperedges (+ attributes)

Inter-object metadata

- **Similarity links** : reflect the strength of the similarity between two objects

Inter-object metadata

- **Similarity links** : reflect the strength of the similarity between two objects



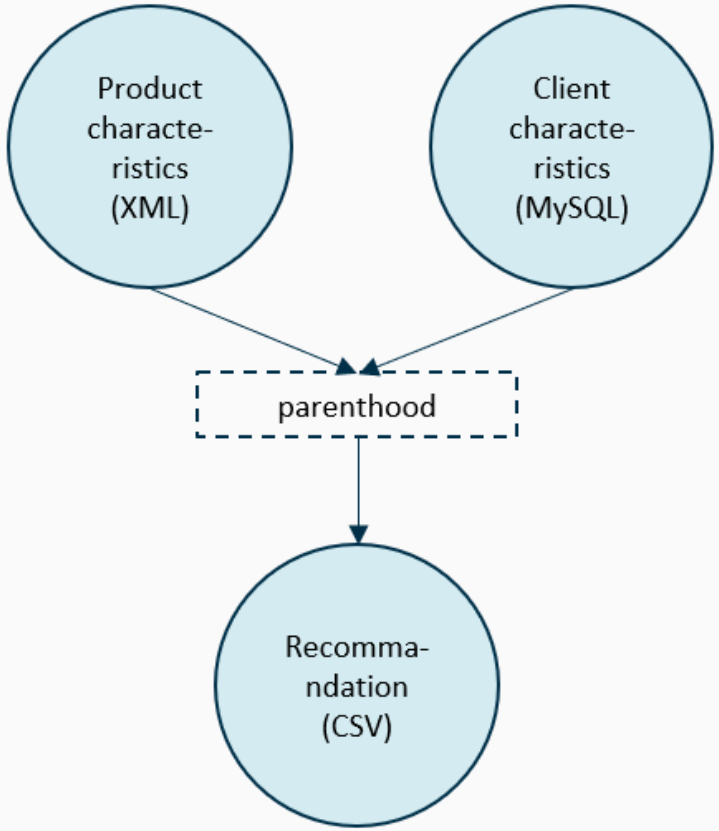
In MEDAL : edges (+ attributes)

Inter-object metadata

- **Parenthood relationships** : object being the result of joining several others

Inter-object metadata

- **Parenthood relationships** : object being the result of joining several others



In MEDAL : directed hyperedges (+ attributes)

Global metadata

- **Semantic ressources** : knowledge bases $\rightarrow$ generate other metadata, improve analyses

Global metadata

- **Semantic ressources** : knowledge bases → generate other metadata, improve analyses
- **Indexes** : data structures → help find objects quickly

Global metadata

- **Semantic ressources** : knowledge bases → generate other metadata, improve analyses
- **Indexes** : data structures → help find objects quickly
- **Logs** : event recording → track data lake usage

Global metadata

- **Semantic ressources** : knowledge bases → generate other metadata, improve analyses
- **Indexes** : data structures → help find objects quickly
- **Logs** : event recording → track data lake usage

Global metadata

- **Semantic ressources** : knowledge bases → generate other metadata, improve analyses
- **Indexes** : data structures → help find objects quickly
- **Logs** : event recording → track data lake usage

No special representation or definition in MEDAL

Conclusion

To summarize

We :

To summarize

We :

- Defined the concept of a data lake precisely
- Identified 6 key features for metadata systems
- Introduced a metadata typology in three categories
- Proposed MEDAL, a graph modeling of the metadata typology

To summarize

We :

- Defined the concept of a data lake precisely
- Identified 6 key features for metadata systems
- Introduced a metadata typology in three categories
- Proposed MEDAL, a graph modeling of the metadata typology

What's next ?

To summarize

We :

- Defined the concept of a data lake precisely
- Identified 6 key features for metadata systems
- Introduced a metadata typology in three categories
- Proposed MEDAL, a graph modeling of the metadata typology

What's next ?

- Implementation(s) of MEDAL, adaptability
- Development of a complete Data Lake
- Tests (scalability), comparisons

Thank you for your attention!

Questions?

References i


Alrehamy, H. and Walker, C. (2015).
Personal Data Lake With Data Gravity Pull.
In *BDCloud 2015, Dalian, china*, volume 88 of *IEEE Computer Society Washington*, pages 160–167.


Ansari, J. W., Karim, N., Decker, S., Cochez, M., and Beyan, O. (2018).
Extending Data Lake Metadata Management by Semantic Profiling.
In *ESWC 2018, Heraklion, Crete, Greece*, ESWC, pages 1–15.


Beheshti, A., Benatallah, B., Nouri, R., Chhieng, V. M., Xiong, H., and Zhao, X. (2017).
CoreDB: a Data Lake Service.
In *CIKM 2017, Singapore, Singapore*, ACM, pages 2451–2454.

References ii


Beheshti, A., Benatallah, B., Nouri, R., and Tabebordbar, A. (2018).
CoreKG: A Knowledge Lake Service.
Proceedings of the VLDB Endowment, 11(12):1942–1945.


Diamantini, C., Giudice, P. L., Musarella, L., Potena, D., Storti, E., and Ursino, D. (2018).
A New Metadata Model to Uniformly Handle Heterogeneous Data Lake Sources.
In *ADBIS 2018 Short Papers and Workshop, Budapest, Hungary*, pages 165–177.


Dixon, J. (2010).
Pentaho, Hadoop, and Data Lakes.
https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

References iii


Fang, H. (2015).
Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem.
In *CYBER 2015, Shenyang, China*, IEEE, pages 820–824.


Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H.-F., and Chu, X. (2016).
CLAMS: Bringing Quality to Data Lakes.
In *SIGMOD 2016, San Francisco, CA, USA*, ACM, pages 2089–2092.


Farrugia, A., Claxton, R., and Thompson, S. (2016).
Towards Social Network Analytics for Understanding and Managing Enterprise Data Lakes.
In *ASONAM 2016, San Francisco, CA, USA*, IEEE, pages 1213–1220.

References iv

Hai, R., Geisler, S., and Quix, C. (2016).
Constance: An Intelligent Data Lake System.
In *SIGMOD 2016, San Francisco, CA, USA*, ACM Digital Library, pages 2097–2100.

Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., and Whang, S. E. (2016).
Managing Google's data lake: an overview of the GOODS system.
In *SIGMOD 2016, San Francisco, CA, USA*, ACM, pages 795–806.

References v

Hellerstein, J. M., Sreekanti, V., Gonzalez, J. E., Dalton, J., Dey, A., Nag, S., Ramachandran, K., Arora, S., Bhattacharyya, A., Das, S., Donsky, M., Fierro, G., She, C., Steinbach, C., Subramanian, V., and Sun, E. (2017).

Ground: A Data Context Service.

In *CIDR 2017, Chaminade, CA, USA*.

Khine, P. P. and Wang, Z. S. (2017).

Data Lake: A New Ideology in Big Data Era.

In *WCSN 2017, Wuhan, China*, volume 17 of *ITM Web of Conferences*, pages 1–6.

Maccioni, A. and Torlone, R. (2018).

KAYAK: A Framework for Just-in-Time Data Preparation in a Data Lake.

In *CAiSE 2018,Tallin, Estonia*, pages 474–489.

References vi


Madera, C. and Laurent, A. (2016).
The next information architecture evolution: the data lake wave.
In *MEDES 2016, Biarritz, France*, pages 174–180.


Miloslavskaya, N. and Tolstoy, A. (2016).
Big Data, Fast Data and Data Lake Concepts.
In *BICA 2016, NY, USA*, volume 88 of *Procedia Computer Science*, pages 1–6.


O'Leary, D. E. (2014).
Embedding AI and Crowdsourcing in the Big Data Lake.
IEEE Intelligent Systems, 29(5):70–73.


Quix, C., Hai, R., and Vatov, I. (2016).
GEMMS: A Generic and Extensible Metadata Management System for Data Lakes.
In *CAiSE 2016, Ljubljana, Slovenia*, pages 129–136.


Singh, K., Paneri, K., Pandey, A., Gupta, G., Sharma, G., Agarwal, P., and Shroff, G. (2016).
Visual Bayesian Fusion to Navigate a Data Lake.
In *FUSION 2016, Heidelberg, Germany*, IEEE, pages 987–994.


Sirosh, J. (2016).
The Intelligent Data Lake.
https://azure.microsoft.com/fr-fr/blog/the-intelligent-data-lake/.

References viii


Suriarachchi, I. and Plale, B. (2016).
Crossing Analytics Systems: A Case for Integrated Provenance in Data Lakes.
In *e-Science 2016, Baltimore, MD, USA*, pages 349–354.


Terrizzano, I., Schwarz, P., Roth, M., and Colino, J. E. (2015).
Data Wrangling: The Challenging Journey from the Wild to the Lake.
In *CIDR 2015, Asilomar, CA, USA*, pages 1–9.

Intra-object metadata definition

Let $\mathcal{N}$ be a set of nodes. The set of *intra-object metadata* $\mathcal{M}_{intra}$ is the set of hypernodes such that $\forall h \in \mathcal{M}_{intra}, h = \langle N, E \rangle$, where :

- $N \subset \mathcal{N}$ is the set of nodes (representations and versions) carrying attributes of h and
- $E = \{r_{(transformation \mid update)} \in N \times N\}$ is the set of edges (transformations and updates) carrying attributes of h.

Intra-object metadata definition

The set of *inter-object metadata* $\mathcal{M}_{inter}$ is defined by three pairs $\langle H, E_g \rangle$, $\langle H', E_s \rangle$ and $\langle H'', E_p \rangle$, where :

- $H \subset \mathcal{M}_{intra}$, $H' \subset \mathcal{M}_{intra}$ and $H'' \subset \mathcal{M}_{intra}$ are sets of hypernodes carrying attributes ;
- $E_g = \{E_g^{param} \mid E_g^{param} : H \rightarrow \mathcal{P}(H)\}$ is the set of functions grouping hypernodes in collections w.r.t. a given parameter (often an attribute) ;
- $E_s = \{s \mid s \in H' \times H'\}$ is the set of edges (similarity links) carrying attributes ;
- $E_p = \{(h_1, ..., h_n, h_{child}) \mid (h_1, ..., h_n, h_{child}) \in (H'')^{n+1}\}$ is the set of parenthood relationships, with $(h_1, ..., h_n)$ being the parent hypernodes ($n \geq 2$) and h_{child} the child hypernode.