Data Lakes: A Survey of Functions and Systems

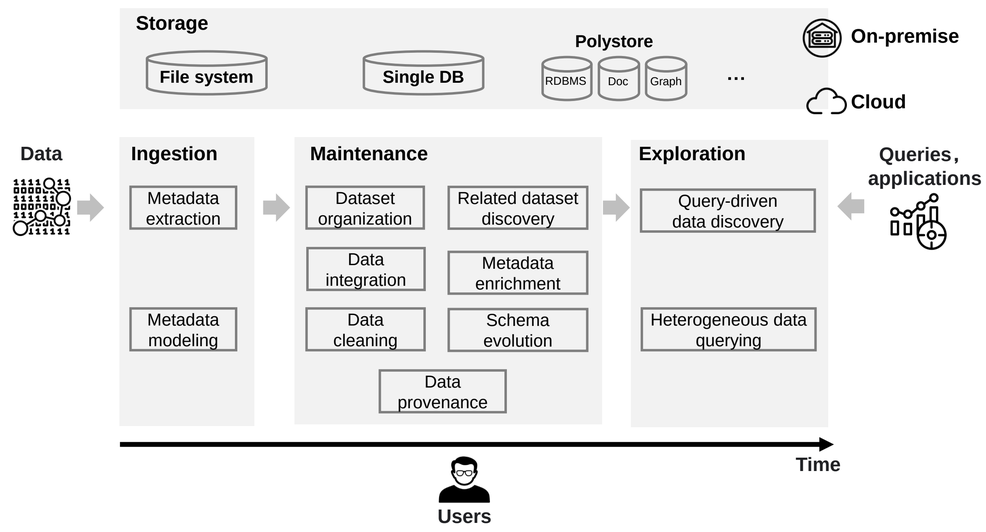
Rihan Hai , Christos Koutras , Christoph Quix , and Matthias Jarke

What is a data lake?

A **data lake** is a flexible, scalable *data storage and management system*, which ingests and stores *raw* data from *heterogeneous sources* in their original format and provides maintenance, query processing, and data analytics in an *on-the-fly* manner, with the help of rich *metadata*.

Storage systems for your data lake: A data lake designer needs to factor in not only the *raw data* but also *how the data will be used*. The choices are diverse: file systems or databases (relational or NoSQL), single or hybrid systems, on-premise or cloud, etc. The specific choice of storage strategy often shapes the required functions.

Architecture



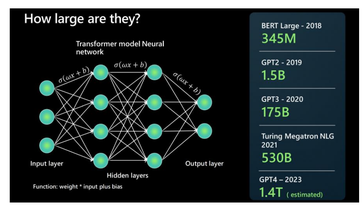
Existing Solutions

Tier	Functions	Systems
Ingestion	Metadata extraction	GEMMS [117]
		DATAMARAN [53]
		Skluma [137]
	Metadata modeling	GEMMS [64], [117]
		HANDLE [43]
		Data vault [57], [107]
		Diamantiniet al. [34], [35], [36]
		Aurum [48]
		Sawadogoet et al. [127]
Maintenance	Dataset organization	GOODS [67], [68]
		DS-Prox [3], [4], [5]
		KAYAK [90], [91]
		Nargesian et al. [104]
		Ronin [110]
		Juneau [152]
	Related dataset discovery	Aurum [48]
		Brackenbury et.al. [15]
		JOSIE [155]
		D^{3}L [14]
		Juneau [75], [151], [152]
		PEXESO [40]
		RNLIM [121]
		DLN [12]
	Data integration	Constance [61], [62], [63], [65]
		ALITE [82]
	Metadata enrichment	CoreDB [9], [10]
		D^4 [109]
		DomainNet [85]
		Constance [64]
		GOODS [67], [68]
	Data cleaning	CLAMS [47]
		Constance [64]
		Song et al. [138]
	Schema evolution	Klettkeet et al. [83]
	Data provenance	IBM tool [143]
		Suriarachchi et al. [141]
		GOODS [67], [68]
		CoreDB [9], [10]
		Juneau [75], [151], [152]
Exploration	Query-driven data discovery	JOSIE [155]
		D^{3}L [14]
		Juneau [75], [151], [152]
		Aurum [48]
	Heterogeneous data querying	Constance [61], [65]
		CoreDB [9], [10]
		Ontario [44], [80]
		Squerall [94]

Future Directions

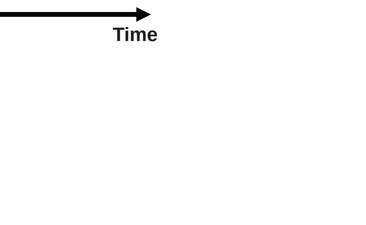
- Data lakes meet machine learning
  - Training data **heterogeneity**
  - **In-lake** machine learning
  - ML workflow **optimization**
  - ML-driven **metadata** management
- Advanced analytics and transaction management
- Data lakes in digital business transformation

ML&DL models



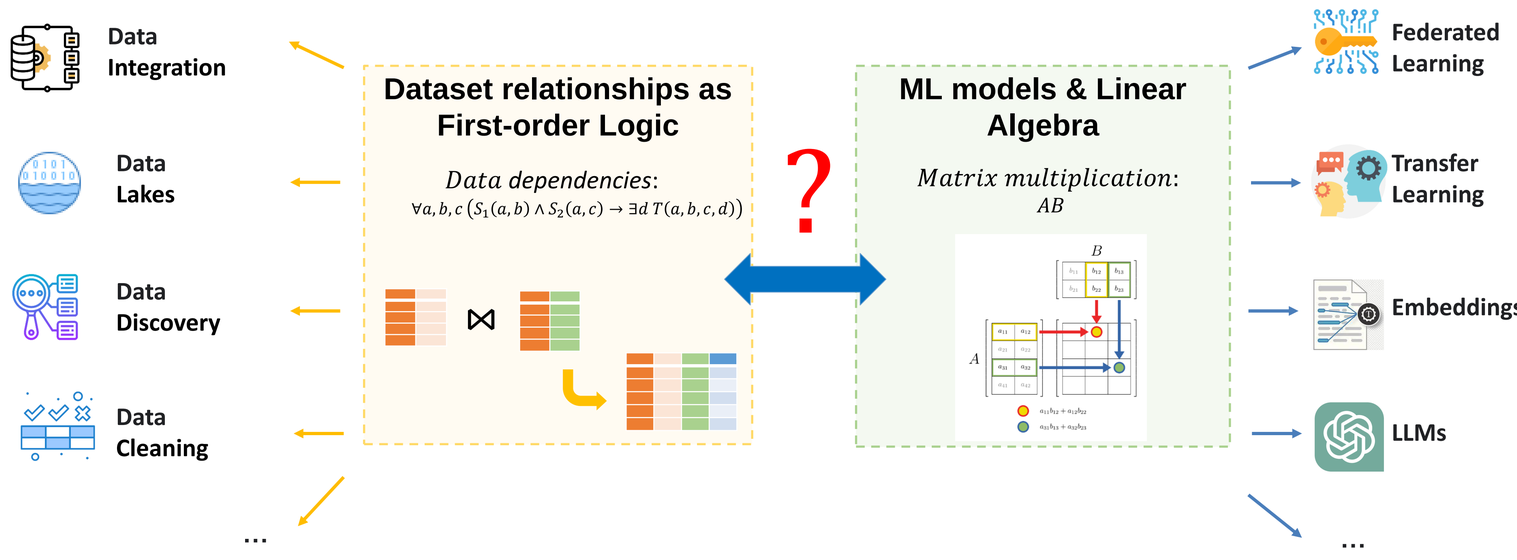
Cloud Computing

New Hardware



Our Vision

Q: Can we use metadata to improve the effectiveness and efficiency of ML model training?



VENI

InfiniData

ExtremeXP

TU Delft Web Information Systems Group

Survey (TKDE'23)

Vision (TKDE'24)