TECHNISCHE UNIVERSITÄT BERLIN

Data Integration and Large-scale Analysis (DIA)
07 Data Provenance and Data Catalogs

Prof. Dr. Matthias Boehm
Technische Universität Berlin
Berlin Institute for the Foundations of Learning and Data
Big Data Engineering (DAMS Lab)

Last update: Nov 28, 2024

BIFOLD

Announcements / Administrative Items

- **#1 Video Recording**
  - Hybrid lectures: in-person H 0107, zoom live streaming, video recording
  - https://tu-berlin.zoom.us/j/9529634787?pwd=R1ZsN1M3SC9BOU1OcFdmem9zT202UT09

- **#2 Exercises/Projects**
  - **Reminder:** exercise/project submissions by **Jan 30** (no extensions)
  - Make use of **virtual** office hours **Wed 4.30pm-6pm**

- **#3 Lecture Dec 05**
  - FONDA II CRC (Foundations of Workflows for Large-Scale Scientific Data Analysis) **retreat Dec 04 – 06**
  - Virtual lecture on **Dec 05, 4pm** (start 4.15pm) from hotel room

Agenda

- **Motivation and Terminology**
- **Data Provenance**
- **Data Catalogs**

Motivation and Terminology

Excursus: FAIR Data Principles

[https://www.go-fair.org/fair-principles/]

- **#1 Findable**
  - Metadata and data have globally unique **persistent identifiers**
  - Data describes w/ rich **meta data**; registered/indexes and searchable
- **#2 Accessible**
  - Metadata and data retrievable via open, free and universal **commnication protocols**
  - Metadata accessible even when data no longer available
- **#3 Interoperable**
  - Metadata and data use a formal, **accessible, and broadly applicable format**
  - Metadata and data use FAIR vocabularies and qualified references
- **#4 Reusable**
  - Metadata and data described with plurality of accurate and relevant attributes
  - Clear license, **associated with provenance**, meets community standards

Terminology of Provenance/Lineage

- **Data Provenance**
  - Track and understand data origins and transformations of data (**where?**, **when?**, **who?**, **why?**, **how?**)

  - Contains meta data, context, and modifications (transform, enrichment)
  - **Synonyms:** **data provenance** (arts) **data lineage** (royals), **data pedigree** (horses)

- **Blockchain**
  - Data structure logging transactions in **verifiable** and **permanent way**

- **Data Catalogs** (curation/**governance**)
  - Directory of datasets including data provenance (meta data, artifacts)
  - Raw/original, curated datasets, derived data products

Application and Goals of Provenance

a) High-Level Goals

- **#1 Versioning and Reproducibility** (analogy experiments)
- **#2 Explainability, Interpretability, Verification**

b) Low-Level Goals

- **#3 Full and Partial Reuse of Intermediates**
- **#4 Incremental Maintenance of MatViews, Models, etc**
- **#5 Tape/log of Executed Operations** → Auto Differentiation
- **#6 Recomputation for Caching / Fault Tolerance**
- **#7 Debugging via Lineage Query Processing**

Data Provenance

Overview Data Provenance

- **Def Data Provenance**
  - Information about the **origin** and **creation process** of data
- Example
  - Debugging suspicious query results

```
SELECT Customer, sum(O.Quantity*P.Price)
  FROM Orders O, Products P
  WHERE O.PID = P.PID
  GROUP BY Customer
```

Customer	Sum
A	7620
B	120
C	130
D	75

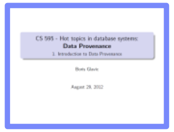
OID	Customer	Date	Quantity	PID
1	A	2019-06-22	3	2
2	B	2019-06-22	1	3
3	A	2019-06-22	101	4
4	C	2019-06-23	2	2
5	D	2019-06-23	1	4
6	C	2019-06-23	1	1

PID	Product	Price
1	X	100
2	Y	15
4	Z	75
3	W	120

Overview Data Provenance, cont.

- **An Abstract View**
  - **Data:** schema, structure → data items
  - **Data composition** (granularity): attribute, tuple, relation
  - **Transformation:** consumes inputs, produces outputs
  - **Hierarchical transformations:** query w/ views, query block, operators
  - **Additional:** env context (OS, libraries, env variables, state), users

[Boris Glavic: CS595 Data Provenance – Introduction to Data Provenance, **Illinois Institute of Technology, 2012**]



- **Goal: Tracing of Derived Results**
  - Inputs and parameters
  - Steps involved in creating the result
  - → Store and query data & provenance
  - General Data Protection Regulation (**GDPR**)?

```
1. Read file1
2. Sort rows
3. Compute median
4. Write to file2
```

Prov.

[Zachary G. Ives: Data Provenance: Challenges, Benefits, Research, **NIH Webinar 2016**]

Classification of Data Provenance

- **Overview**
  - Base query $Q(D) = O$ with database $D = \{R_1, \ldots, R_n\}$
  - **Forward lineage query:** $L_f(R_i'', O')$ from subset of input relation to output
  - **Backward lineage query:** $L_b(O', R_i)$ from subset of outputs to base tables

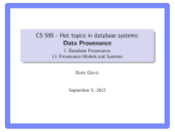
- **#1 Lazy Lineage Query Evaluation**
  - Rewrite (**invert**) lineage queries as relational queries over input relations
  - No runtime overhead but slow lineage query processing

- **#2 Eager Lineage Query Evaluation**
  - Materialize **annotations** (data/transforms) during base query evaluation
  - Runtime overhead but fast lineage query processing
  - Lineage capture: **Logical** (relational) vs **physical** (instrumented physical ops)

[Fotis Psallidas, Eugene Wu: Smoke: Fine-grained Lineage at Interactive Speed. **PVLDB 2018**]

Why-Provenance

[Boris Glavic: CS595 Data Provenance – Provenance Models and Systems, **Illinois Institute of Technology, 2012**]



- **Overview Why**
  - **Goal:** Which input tuples contributed to an output tuple t in query Q
  - Representation: Set of witnesses w for tuple t (**set semantics!**)
    - $w \subseteq I$ (subset of all tuples in instance I)
    - $t \in Q(w)$ (tuple in result of query over w)

- **Example Witnesses**

```
SELECT Customer, Product
  FROM Orders O, Products P
  WHERE O.PID=P.PID
```

	Customer	Date	PID
o1	A	2019-06-22	2
o2	B	2019-06-22	3
o3	A	2019-06-22	2

	PID	Product
p1	1	X
p2	2	Y
p3	4	Z
p4	3	W

	Customer	Product
t1	A	Y
t2	B	W

Witnesses for **t1**:
w1 = {o1,p2}, w2 = {o3,p2},
w3 = {o1,o3,p2}, …, wn = I

Minimal witnesses for t1:
w1 = {o1,p2}, w2 = {o3,p2}

How-Provenance

[Boris Glavic: CS595 Data Provenance – Provenance Models and Systems, **Illinois Institute of Technology, 2012**]

- **Overview**
  - Model how tuples where combined in the computation
  - **Alternative use:** need one of the tuples (e.g., union/projection)
  - **Conjunctive use:** need all tuples together (e.g., join)
- **Provenance Polynomials**
  - **Semi-ring annotations** to model provenance $(\mathbb{N}[I], +, \times, 0, 1)$
- **Examples**
  - $q = \pi_a(R)$

	a	b
r1	1	2
r2	1	3

a	Provenance
1	r1 + r2

  - $q = \pi_b(R \bowtie S)$

	a	b
r1	1	P
r2	2	G
	3	M

	c	a
s1	S	1
s2	S	2
s3	W	2

b	Provenance
P	r1 x s1
G	(r2 x s2) + (r2 x s3)

Provenance Polynomials

How-Provenance, cont.

- **Example Exam Question:**
  Given below tables R and S (with tuples r_i and s_j), query Q and results O, specify the provenance polynomials for every tuple in O. [3 points]

R

	A	B
r_1	X	1
r_2	Y	2
r_3	Z	1

S

	C	D
s_1	1	A
s_2	2	B
s_3	2	A
s_4	2	C

```
SELECT DISTINCT S.D
  FROM R, S
  WHERE R.B=S.C
```

⇒

O	Provenance Polynomials?
A	
B	
C	

A: r1 x s1 + r3 x s1 + r2 x s3
(equivalent: **(r1 + r3) x s1 + r2 x s3**)

B: r2 x s2

C: r2 x s4

Why Not?-Provenance

[Adriane Chapman, H. V. Jagadish: **Why not? SIGMOD 2009**]

- **Overview**
  - Why are items not in the results
  - **Example Problem:**
    "Window-display-books < $20"
    → (Euripides, Medea).
    → **Why not** (Hrotsvit, Basilius)?

<= 20$?

Not in book store?

Bug in the query / system?

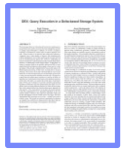
Author	Title	Price	Publisher
	Epic of Gilgamesh	$150	Hesperus
Euripides	Medea	$16	Free Press
Homer	Iliad	$18	Penguin
Homer	Odyssey	$49	Vintage
Hrotsvit	Basilius	$20	Harper
Longfellow	Wreck of the Hesperus	$89	Penguin
Shakespeare	Coriolanus	$70	Penguin
Sophocles	Antigone	$48	Free Press
Virgil	Aeneid	$92	Vintage

- **Evaluation Strategies**
  - Given a user question (why no tuple satisfies predicate S), dataset D, result set R, and query Q, leverage **why lineage**
  - **#1 Bottom-Up:** from leafs in topological order to find last op eliminating $d \in S$
  - **#2 Top-Down:** from result top down to find last op, **requires stored lineage**

Provenance for ML Pipelines (fine-grained)

- **DEX: Dataset Versioning**
  - **Versioning of datasets, stored with delta encoding**
  - Checkout, intersection, union queries over deltas
  - Query optimization for finding efficient plans

[Amit Chavan, Amol Deshpande: DEX: Query Execution in a Delta-based Storage System. **SIGMOD 2017**]



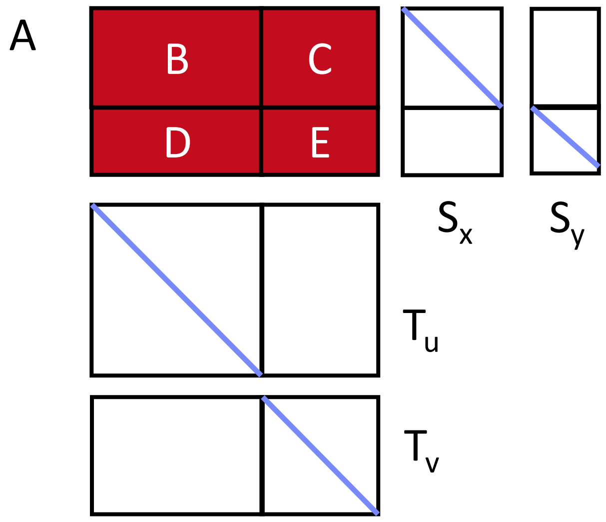
- **MISTIQUE: Intermediates of ML Pipelines**
  - Capturing, storage, querying of intermediates
  - **Lossy deduplication and compression**
  - Adaptive querying/materialization for finding efficient plans

[Manasi Vartak et al: MISTIQUE: A System to Store and Query Model Intermediates for Model Diagnosis. **SIGMOD 2018**]

- **Linear Algebra Provenance**
  - Provenance propagation by decomposition
  - Annotate parts w/ provenance polynomials (contributing inputs + impact)

[Zhepeng Yan, Val Tannen, Zachary G. Ives: Fine-grained Provenance for Linear Algebra Operators. **TaPP 2016**]

$$A = S_x B T_u + S_x C T_v$$
$$+ S_y D T_u + S_y E T_v$$



Provenance for ML Pipelines (coarse-grained)

- **MLflow**
  - Programmatic API for tracking parameters, experiments, and results
  - **`autolog()`** for specific params

[**Credit:** https://databricks.com/blog/2018/06/05]

```
import mlflow
mlflow.log_param("num_dimensions", 8)
mlflow.log_param("regularization", 0.1)
mlflow.log_metric("accuracy", 0.1)
mlflow.log_artifact("roc.png")
```

- **Flor (on Ground)**
  - DSL embedded in python for managing the workflow development phase of the ML lifecycle
  - DAGs of actions, artifacts, and literals
  - Data context generated by activities in Ground

https://rise.cs.berkeley.edu/projects/jarvis/

[Joseph M. Hellerstein et al: Ground: A Data Context Service. **CIDR 2017**]

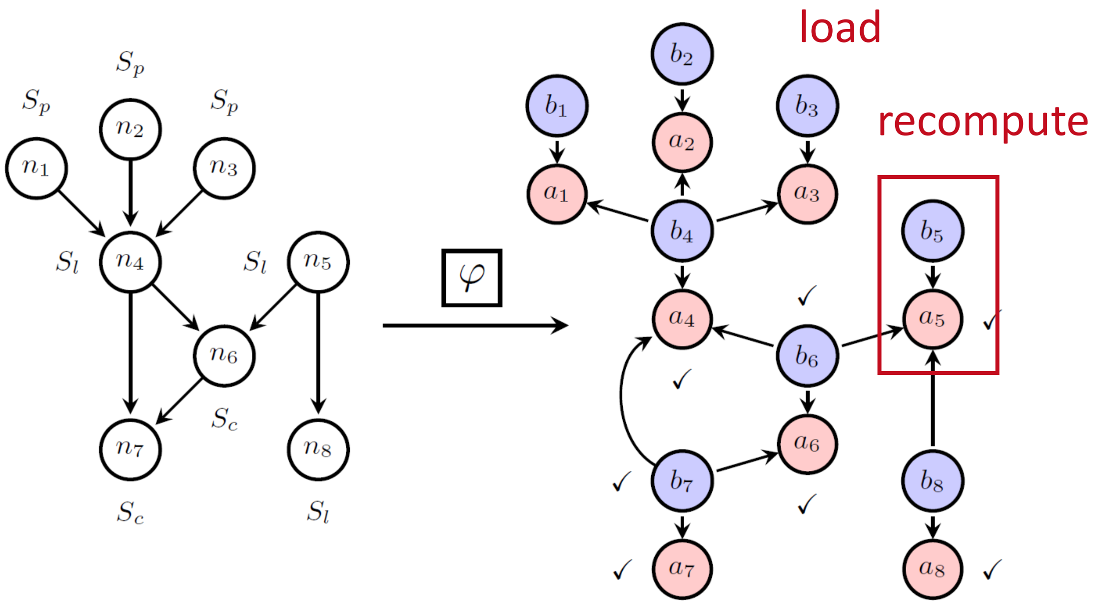
- **Vision: Dataset Relationship Management**
  - **Reuse**, **reveal**, **revise**, **retarget**, **reward**
  - Code-to-data relationships (data provenance)
  - Data-to-code relationships (potential transforms)

[Zachary G. Ives, Yi Zhang, Soonbo Han, Nan Zheng,: Dataset Relationship Management. **CIDR 2019**]

Provenance for ML Pipelines (coarse-grained), cont.

- **HELIX**
  - Goal: focus on iterative development w/ small modifications (trial & error)
  - Caching, reuse, and recomputation
  - Reuse as **Max-Flow problem** → **NP-hard** → heuristics
  - Materialization to disk for future reuse

[Doris Xin, Stephen Macke, Litian Ma, Jialin Liu, Shuchen Song, Aditya G. Parameswaran: Helix: Holistic Optimization for Accelerating Iterative Machine Learning. **PVLDB 2018**]



- **Collaborative Optimizer**

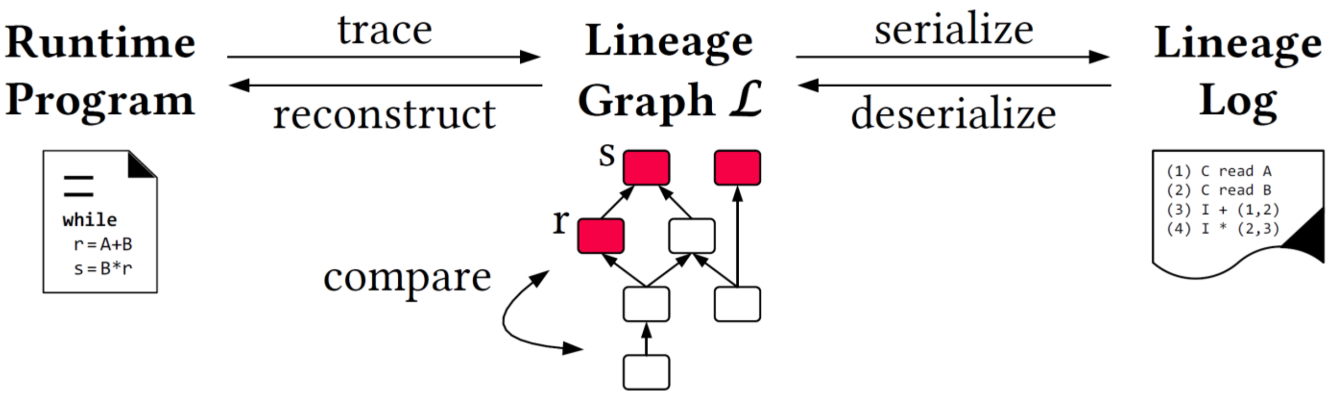
[Behrouz Derakhshan, Alireza Rezaei Mahdiraji, Ziawasch Abedjan, Tilmann Rabl, Volker Markl: Optimizing Machine Learning Workloads in Collaborative Environments. **SIGMOD 2020**]

Lineage Tracing & Reuse in SystemDS [CIDR’20, SIGMOD’21a, EDBT’25]

- **Problem**
  - **Exploratory data science** (data preprocessing, model configurations)
  - **Reproducibility** and **explainability** of trained models (data, parameters, prep)
  - ➔ **Lineage/Provenance as Key Enabling Technique:** Model versioning, reuse of intermediates, incremental maintenance, auto differentiation, and debugging (query processing over lineage)

[Arnab Phani, Benjamin Rath, Matthias Boehm: LIMA: Fine-grained Lineage Tracing and Reuse in Machine Learning Systems, **SIGMOD 2021**]

- **Efficient Lineage Tracing**
  - Tracing of inputs, literals, and **non-determinism**
  - **Trace lineage of logical operations**
  - **Deduplication** for loops/functions
  - Program/output reconstruction



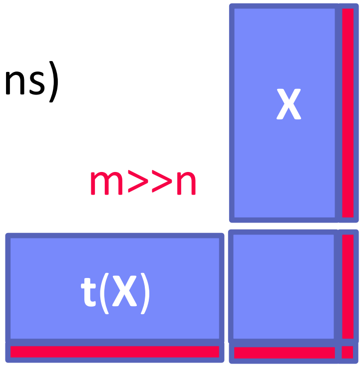
Lineage Tracing & Reuse in SystemDS, cont.


- **Multi-level, Lineage-based Reuse**
  - Lineage trace uniquely identifies intermediates
  - Reuse intermediates at function / block / operation level
- **Full Reuse of Intermediates**
  - Before executing instruction, probe output lineage in cache `Map<Lineage, MatrixBlock>`
  - Cost-based/heuristic caching and eviction decisions (compiler-assisted)
- **Partial Reuse of Intermediates**
  - **Problem:** Often partial result overlap
  - Reuse partial results via dedicated rewrites (compensation plans)
  - Example: steplm
- **Next Steps:** multi-backend, unified mem mgmt

```
for( i in 1:numModels )
  R[,i] = lm(X, y, lambda[i,], ...)
```

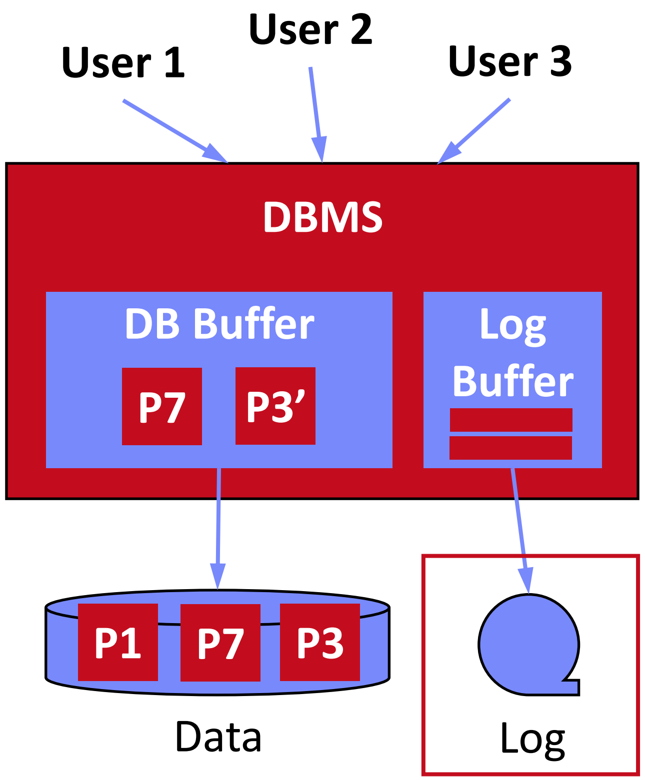
```
m_lmDS = function(...) {
  l = matrix(reg,ncol(X),1)
  A = t(X) %*% X + diag(l)
  b = t(X) %*% y
  beta = solve(A, b) ...}
```

```
m_steplm = function(...) {
  while( continue ) {
    parfor( i in 1:n ) {
      if( !fixed[1,i] ) {
        Xi = cbind(Xg, X[,i])
        B[,i] = lm(Xi, y, ...)
    } }
    # add best to Xg (AIC)
} }
```



Recap: Database (Transaction) Log

- **Database Architecture**
  - **Page-oriented storage** on disk and in memory (DB buffer)
  - Dedicated **eviction algorithms**
  - Modified in-memory pages marked as dirty, flushed by cleaner thread
  - **Log:** append-only TX changes
  - Data/log often placed on different devices and periodically archived (backup + truncate)



- **Write-Ahead Logging (WAL)**
  - The log records of changes to some (dirty) data page must be on **stable storage before the data page** (UNDO - atomicity)
  - **Force-log on commit** or full buffer (REDO - durability)
  - **Recovery:** forward (REDO) and backward (UNDO) processing
  - Log sequence number (LSN)

[C. Mohan, Donald J. Haderle, Bruce G. Lindsay, Hamid Pirahesh, Peter M. Schwarz: ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging. **TODS 1992**]

Bitcoin and Blockchain Fundamentals

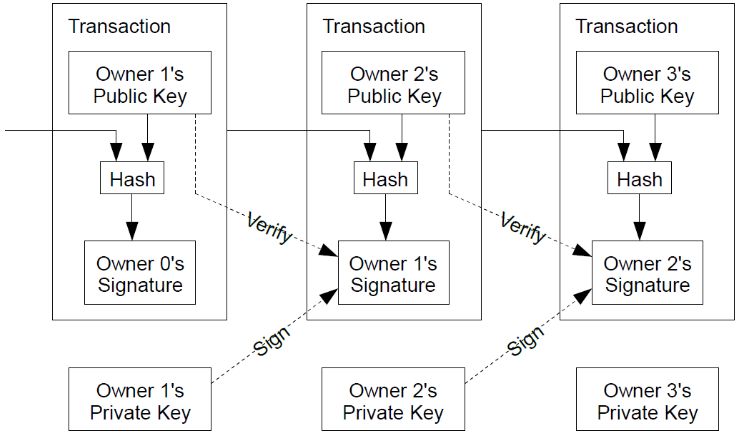
[**Satoshi Nakamoto**: Bitcoin: A Peer-to-Peer Electronic Cash System, **White paper 2008**]

- **Motivation**
  - Peer-to-peer (decentralized, anonymous) electronic cash/payments
  - **Non-reversible transactions** w/o need for trusted third party
  - **Statistics**

	Market Price (USD)	Average Block Size	Transactions per Day	Mempool Size
Nov 21 2019:	$7,862.72 USD	1.16 Megabytes	303,921 Transactions	11,304,890 Bytes
Nov 19 2020:	$17,975.24 USD	1.29 Megabytes	310,424 Transactions	19,920,773 Bytes

[https://www.blockchain.com/charts]

- **Transaction Overview**
  - Electronic coin defined as chain of digital signatures
  - Transfer by signing hash of previous TX and public key of next owner
  - **Double-spending problem** (without global verification)

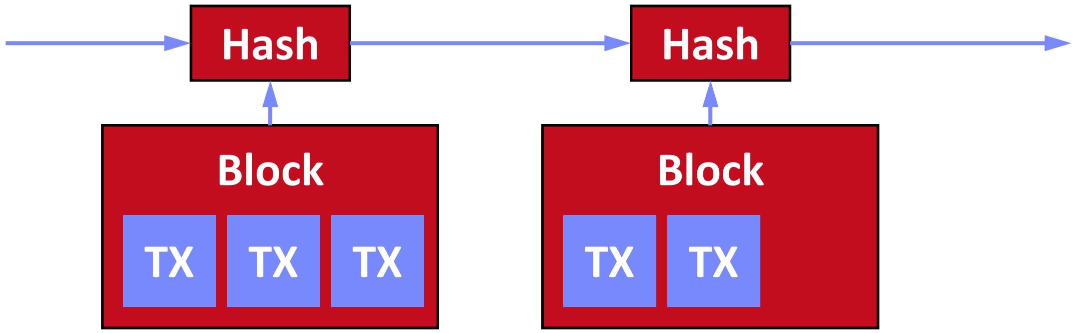


- **Permissioned/Private Blockchains**
  - Blockchain as shared, replicated, permissioned ledger (TX log): **consensus**, **provenance**, **immutability**

Blockchain Data Structure

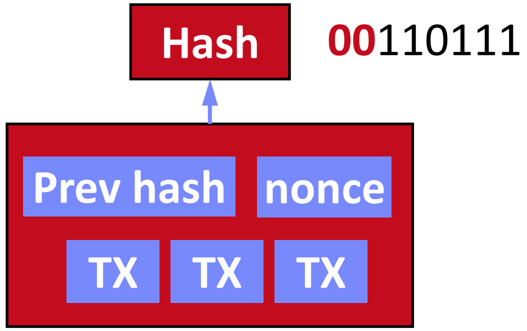
[**Satoshi Nakamoto**: Bitcoin: A Peer-to-Peer Electronic Cash System, **White paper 2008**]

- **Timestamp Server**
  - Decentralized timestamp server: chain of hashes ➔ **public ledger**



Enforces order dependency
➔ **No double-spending**

- **Proof-of-Work**
  - Scanning for value (nonce) whose **SHA-256 hash** begins with a number of zero bits → exponential in number of zeros
  - # zero bits determined by moving average of avg blocks/hour
  - Hard to recompute for chain, easy to check
  - **Majority decision:** CPU time, longest chain



Merkel tree (hash tree)

Blockchain Data Structure, cont.

- **Bitcoin Mining**
  - HW: from CPU to GPUs/FPGAs/ASICs (**10-70 TH/s** @ 2-3KW)
  - Usually mining pools → **"mining cartels"**
- **Hash Rate of Bitcoin Network**
  - **~10 min per block** (144 blocks per day)

@Malaysia

Nov 12, 2021: ~160 EH
Nov 25, 2023: ~494 EH
Nov 25, 2024: ~776 EH

[https://www.blockchain.com/en/charts/hash-rate?daysAverageString=7&timespan=180days, **Nov 12 2021**]

Blockchain Consensus Mechanisms

"means of showing that one invested a non-trivial amount of effort related to some statement"

- **Proof of Work (PoW)**
  - Validation by performing work, existence of HW resources
  - **High HW cost of attacks**
  - **Wasted work, resources, energy** (only first block, no real outcome, e-waste)

- **Proof of Stake (PoS)**
  - Validation by stake-weighted random node selection
  - **Intrinsic coin cost**, **less HW resources**/energy
  - **Untested attack mitigation?**

Ethereum 2.0
→ PoS/sharding over time

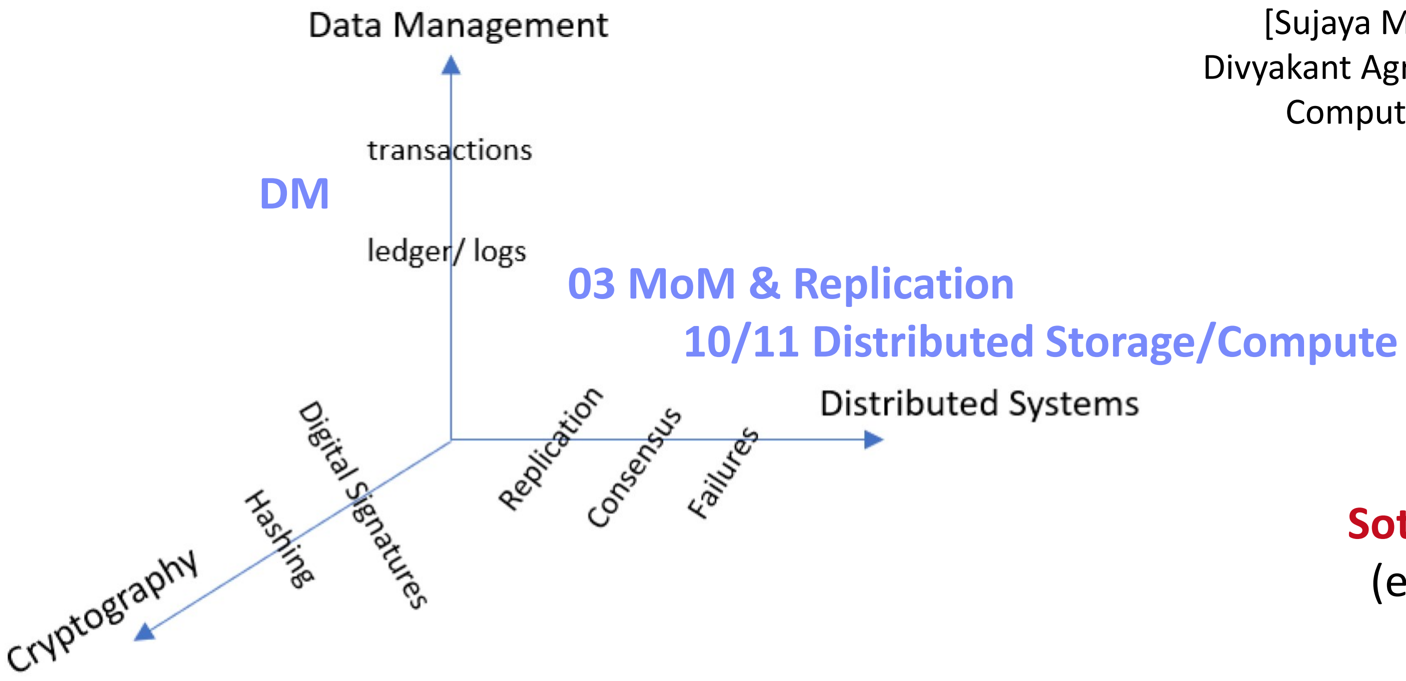
- **Proof of Space/Capacity**
  - Upfront creation of "plot files", store nonces+hashes, find solutions, occasional validation
  - **HW costs of attacks**, use of **unused space**
  - **Moderate adoption**

[https://www.chia.net/]

[Stefan Dziembowski, Sebastian Faust, Vladimir Kolmogorov, Krzysztof Pietrzak: Proofs of Space. IACR Cryptol. 2013]

Discussion Blockchain

[Sujaya Maiyya, Victor Zakhary, Mohammad Javad Amiri, Divyakant Agrawal, Amr El Abbadi: Database and Distributed Computing Foundations of Blockchains. **SIGMOD 2019**]



Many established techniques

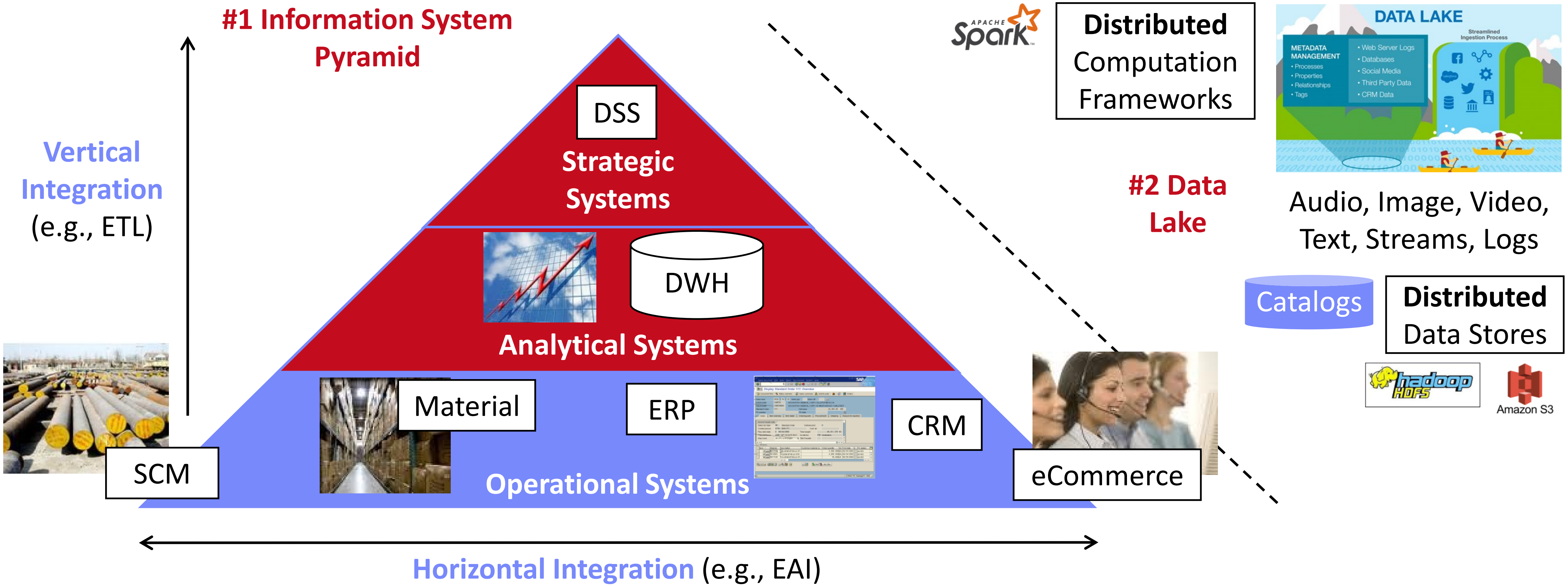
SotA toward scalable/efficient blockchains

(especially for permissioned blockchains)

➔ **Recommendation: Investigate business requirements/context, decide on technical properties and acceptable trade-offs**

Data Catalogs

Recap: Complementary System Architectures



Overview Data Catalogs

- **Data Catalogs**
  - Data curation in repositories for finding datasets in **data lakes**
  - **Metadata and provenance**
  - Augment data with open and linked data sources
- **Examples**

[Alon Y. Halevy et al: Goods: Organizing Google's Datasets. **SIGMOD 2016**]

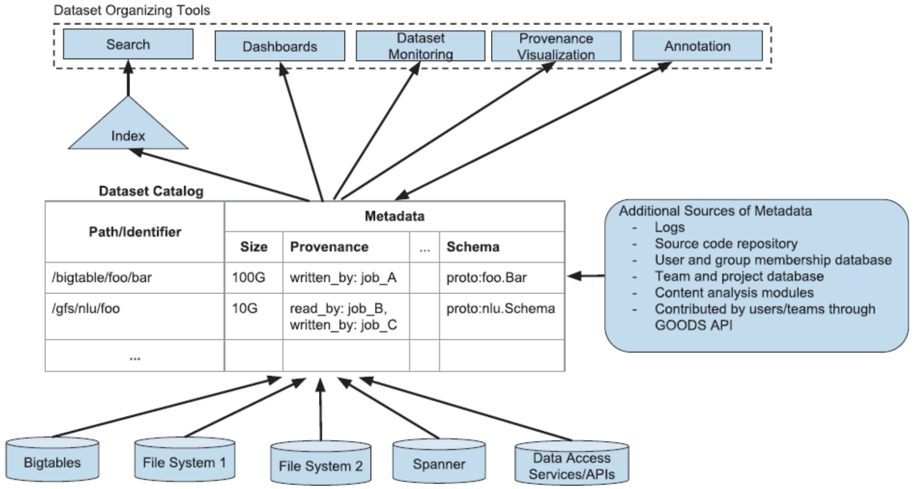
[Dan Brickley, Matthew Burgess, Natasha F. Noy: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. **WWW 2019**]

[Omar Benjelloun, Shiyu Chen, Natasha Noy: Google Dataset Search by the Numbers, https://arxiv.org/pdf/2006.06894]

SAP Data Hub

[SAP Sapphire Now 2019]

Google Dataset Search



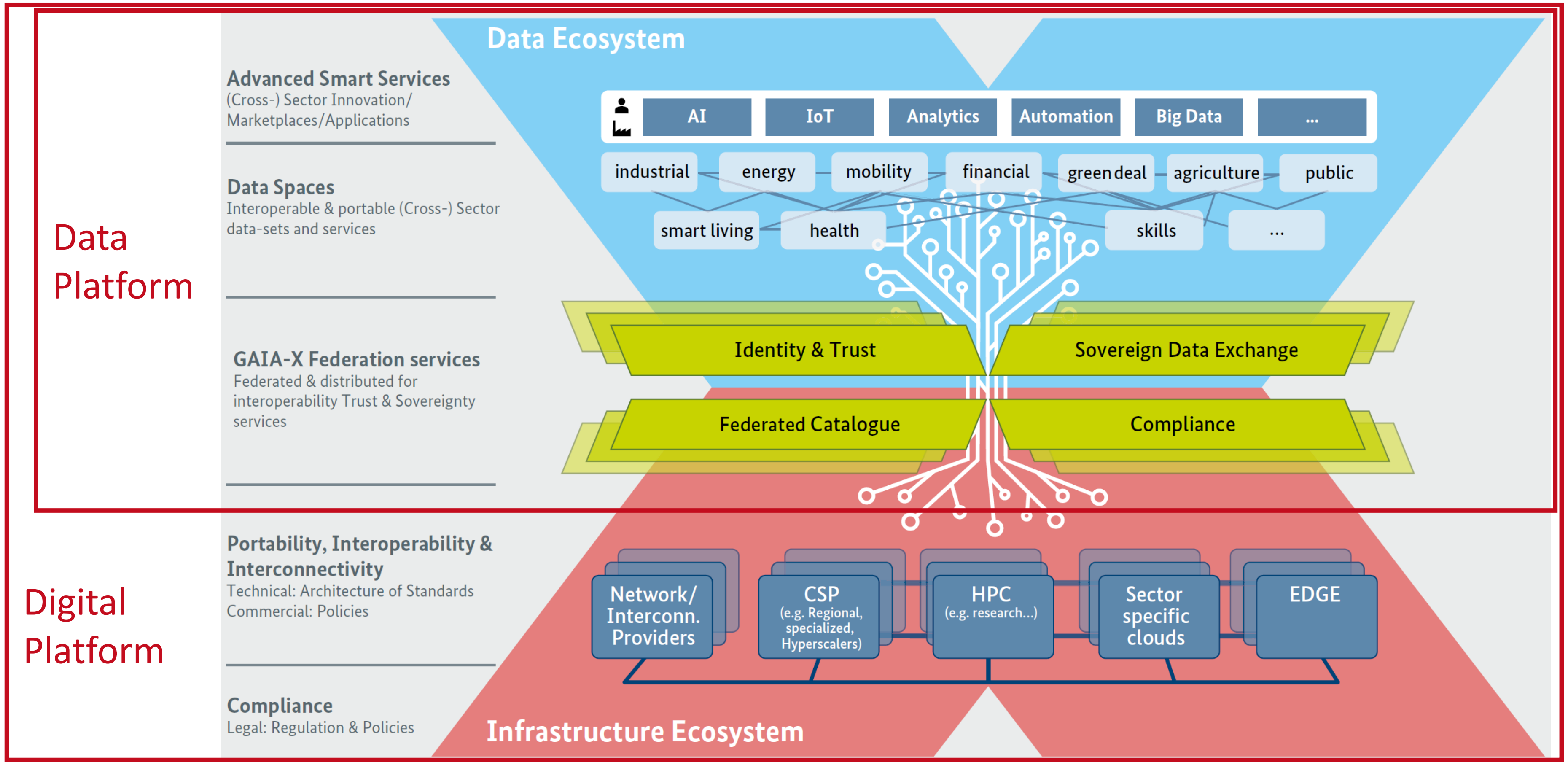
Category	Number of datasets	% of total	Sample formats
Tables	7,822K	37%	CSV, XLS
Structured	6,312K	30%	JSON, XML, OWL, RDF
Documents	2,277K	11%	PDF, DOC, HTML
Images	1,027K	5%	JPEG, PNG, TIFF
Archives	659K	3%	ZIP, TAR, RAR
Text	623K	3%	TXT, ASCII
Geospatial	376K	2%	SHP, GEOJSON, KML
Computational biology	110K	<1%	SBML, BIOPAX2, SBGN
Audio	27K	<1%	WAV, MP3, OGG
Video	9K	<1%	AVI, MPG
Presentations	7K	<1%	PPTX
Medical imaging	4K	<1%	NII, DCM
Other categories	2,245K	11%	

500K → 30M datasets

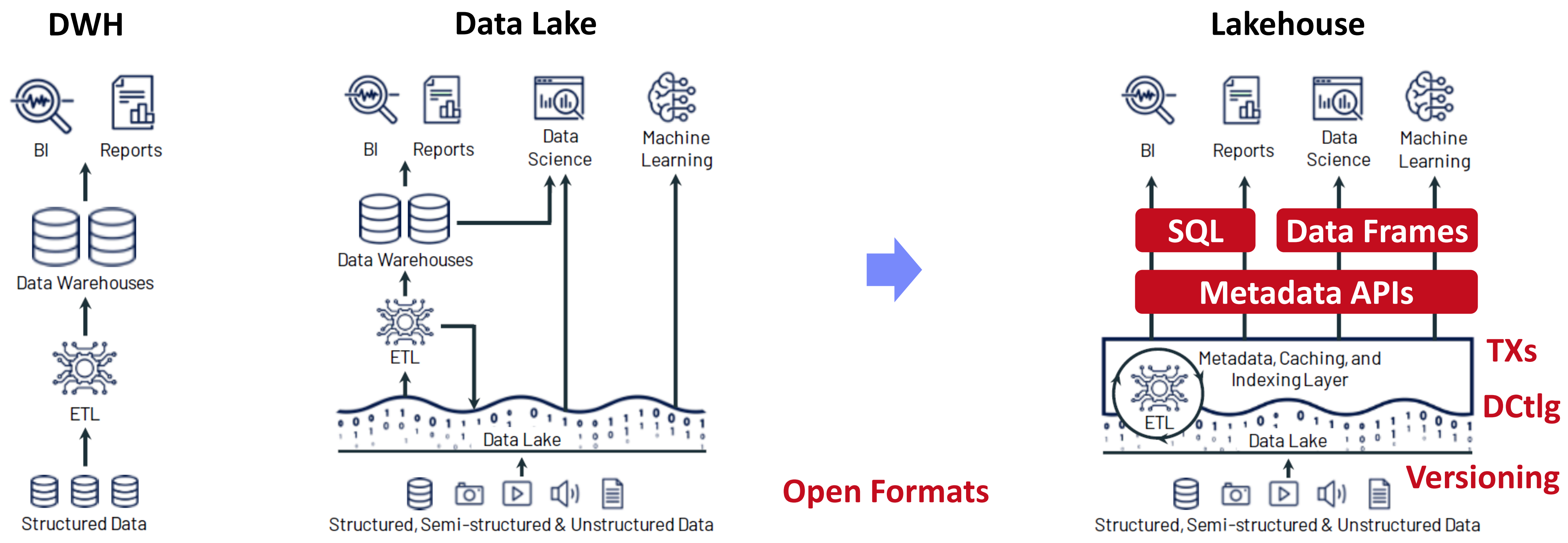
GAIA-X Initiative & Integration

[BMWi: GAIA-X: Driver of digital innovation in Europe – Featuring the next generation of data infrastructure, **2020**]

- **GAIA-X Architecture Overview**



Example Delta Lake (and Lakehouse Architecture)



[Michael Armbrust et al: **Delta Lake:** High-Performance ACID Table Storage over Cloud Object Stores. **PVLDB 13(12) 2020**]

[Michael Armbrust, Ali Ghodsi, Reynold Xin, Matei Zaharia: **Lakehouse:** A New Generation of Open Platforms that Unify Data Ware-housing and Advanced Analytics, **CIDR 2021**]

[Alexander Behm: Photon: A High-Performance Query Engine for the **Lakehouse**, **CIDR 2022**]

Key Features of a Data Catalog

- **#1 Dictionary of Datasets**
  - Basic overview, links, and curation of available datasets
  - Raw/original, curated datasets, derived data products
- **#2 Rich Meta Data Collection**
  - **Format, schema, and access information** of datasets
  - **Data profiling, data validation** results, and data quality scores
- **#3 Lineage/Provenance**
  - Coarse- **or fine-grained lineage**, incl applied **data integration and cleaning process**
  - Optionally **artifacts** to reproduce datasets from sources
- **#4 Data Discovery, Governance, and Sharing**
  - Find related "joinable" datasets (e.g., over spatial-temporal keys)
  - **Efficient discovery** and **sharing** of federated data sources

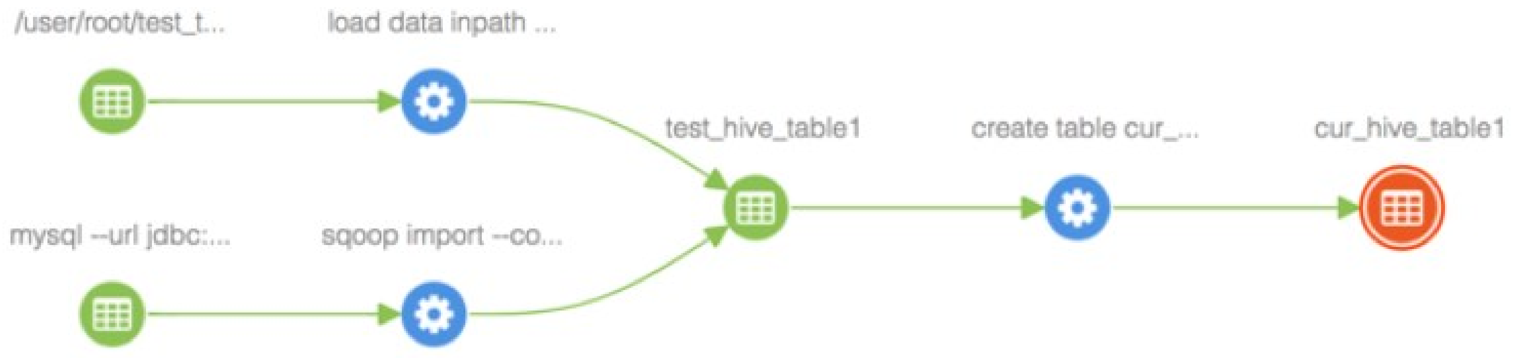
[Sonia Castelo, Rémi Rampin, Aécio S. R. Santos, Aline Bessa, Fernando Chirigati, Juliana Freire: Auctus: A Dataset Search Engine for Data Discovery and Augmentation. **PVLDB 2021**]

Apache Atlas

[https://atlas.apache.org]

- **Apache Atlas Overview**
  - Metadata management and governance capabilities
  - Build catalog (data classification, cross-component lineage)

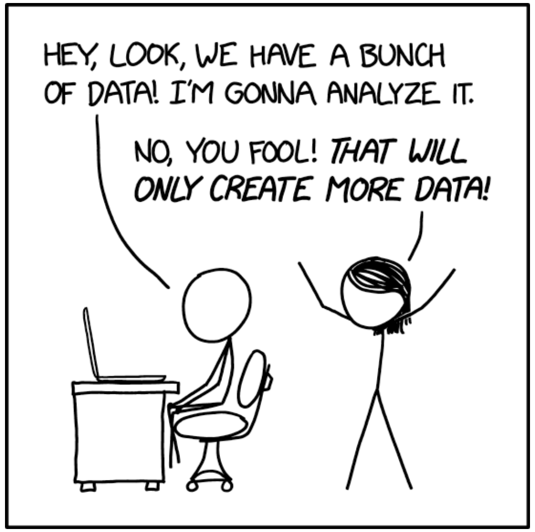
- **Example**
  - Configure Atlas hooks w/ Hadoop components
  - Automatic tracking of lineage and side effects



[https://www.cloudera.com/tutorials/cross-component-lineage-with-apache-atlas-across-apache-sqoop-hive-kafka-storm/.html]

Summary and Q&A

- **Motivation and Terminology**
- **Data Provenance**
- **Data Catalogs**



[https://xkcd.com/2582/]

- **Next Lectures (Large-scale Data Management and Analysis)**
  - **08 Cloud Computing Fundamentals** [Dec 05, **virtual only**]
  - **09 Cloud Resource Management and Scheduling** [Dec 12]
  - **10 Distributed Data Storage** [Dec 19]
  - **11 Distributed, Data-Parallel Computation** [Jan 09]
  - **12 Distributed Stream Processing** [Jan 16]
  - **13 Distributed Machine Learning Systems** [Jan 23]