The metadata challenge

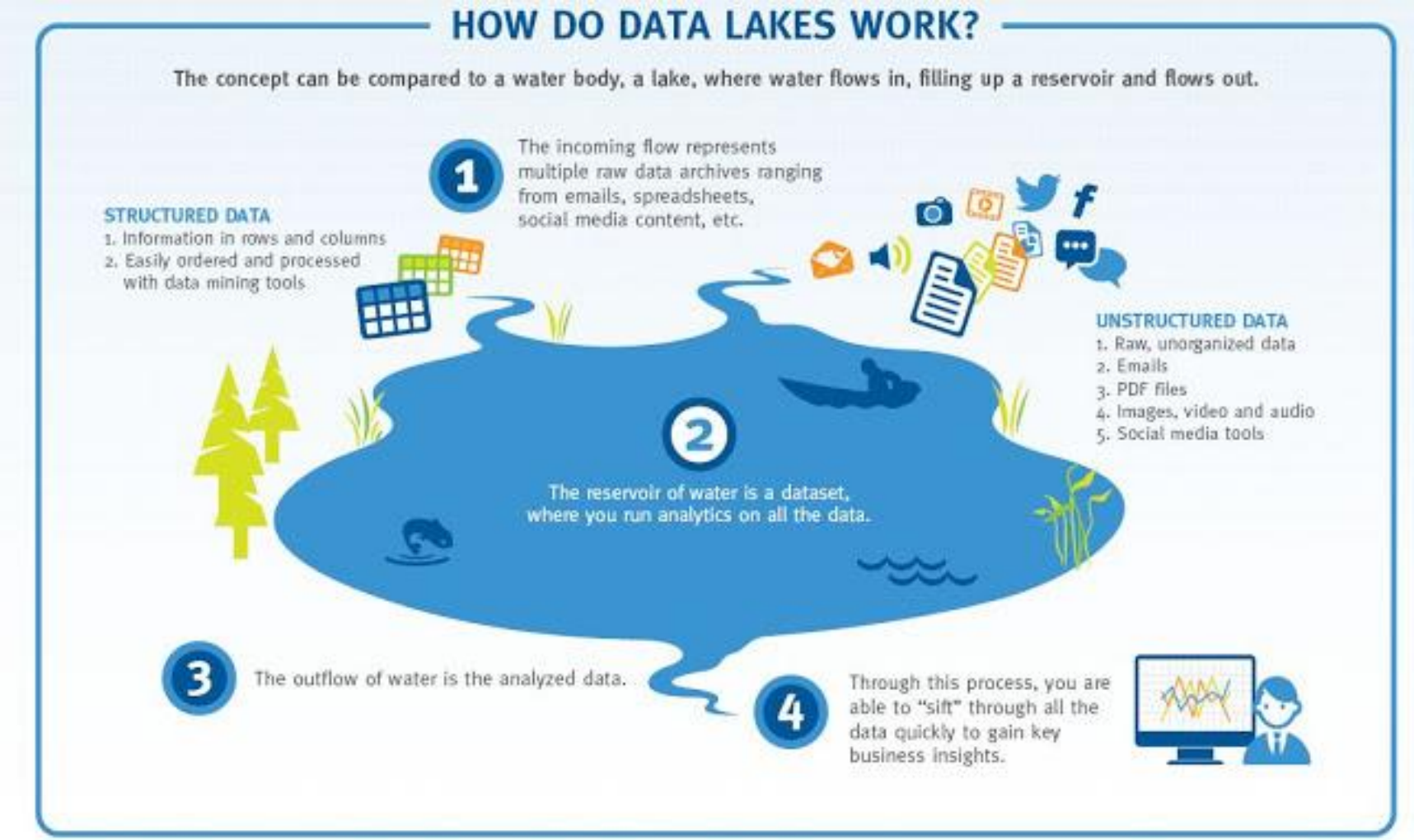
What we are going to do

Definitions: from the data lake to the data platform

Define challenges

Discuss current solutions and future directions

Data lake


HOW DO DATA LAKES WORK?
The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.
1
The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.
STRUCTURED DATA
1. Information in rows and columns
2. Easily ordered and processed with data mining tools
UNSTRUCTURED DATA
1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools
2
The reservoir of water is a dataset, where you run analytics on all the data.
3
The outflow of water is the analyzed data.
4
Through this process, you are able to "sift" through all the data quickly to gain key business insights.

Data lake

"*If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state.*"

- James Dixon, 2010

"*A large storage system for raw, heterogeneous data, fed by multiple data sources, and that allows users to explore, extract and analyze the data.*"

- Sawadogo, P., Darmont, J. **On data lake architectures and metadata management**. *J Intell Inf Syst* 56, 97–120 (2021)

"*A data lake is a central location that holds a large amount of data in its native, raw format.*"

- Databricks, 2021

Data lake

The data lake started with the Apache Hadoop movement, using the Hadoop File System (HDFS) for cheap storage

- *Schema-on-read* architecture
- Agility of storing any data at low cost
- Eludes the problems of quality and governance

A two-tier data lake + warehouse architecture is dominant in the industry

- HDFS replaced by cloud data lakes (e.g., S3, ADLS, GCS)
- Data lake data directly accessible to a wide range of analytics engines
- A subset of data is "ETL-ed" to a data warehouse for important decision support and BI apps

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). **Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics**. *CIDR*.

Data lake

Downsides of data lakes

- Security
  - All the data is stored and managed as files
  - No fine-grained access control on the contents of files, but only coarse-grained access governing who can access what files or directories
- Quality
  - Hard to prevent data corruption and manage schema changes
  - Challenging to ensure atomic operations when writing a group of files
  - No roll-back mechanism
- Query performance
  - Formats are not optimized for fast access

It is often said that the *lake* easily turns into a *swamp*

Data lakehouse

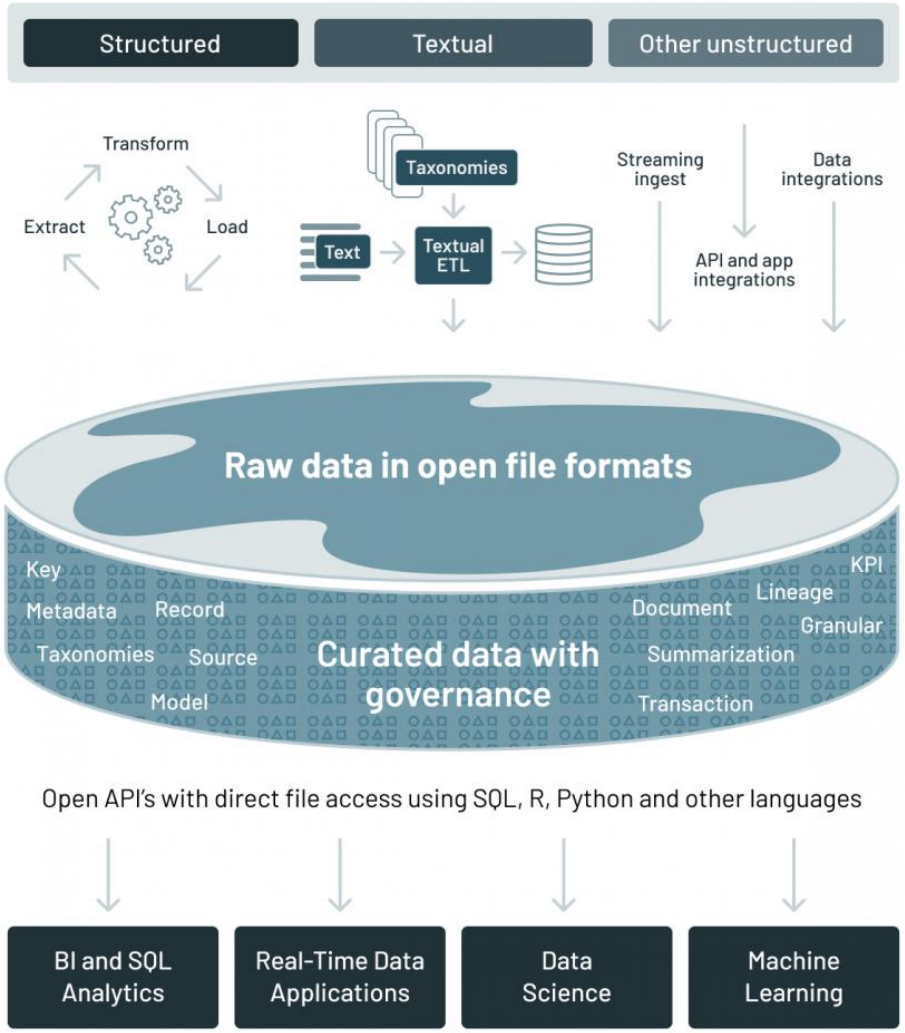
The data lakehouse enables storing all your data once in a data lake and efficiently doing AI and BI on that data directly at a massive scale

- ACID transaction support
- Schema enforcement
- Data governance
  - All processes ensuring that data meet high quality standards throughout the whole lifecycles
  - Including availability, usability, consistency, integrity, security
- Support for diverse workloads (e.g., data science, ML, SQL, analytics)

https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

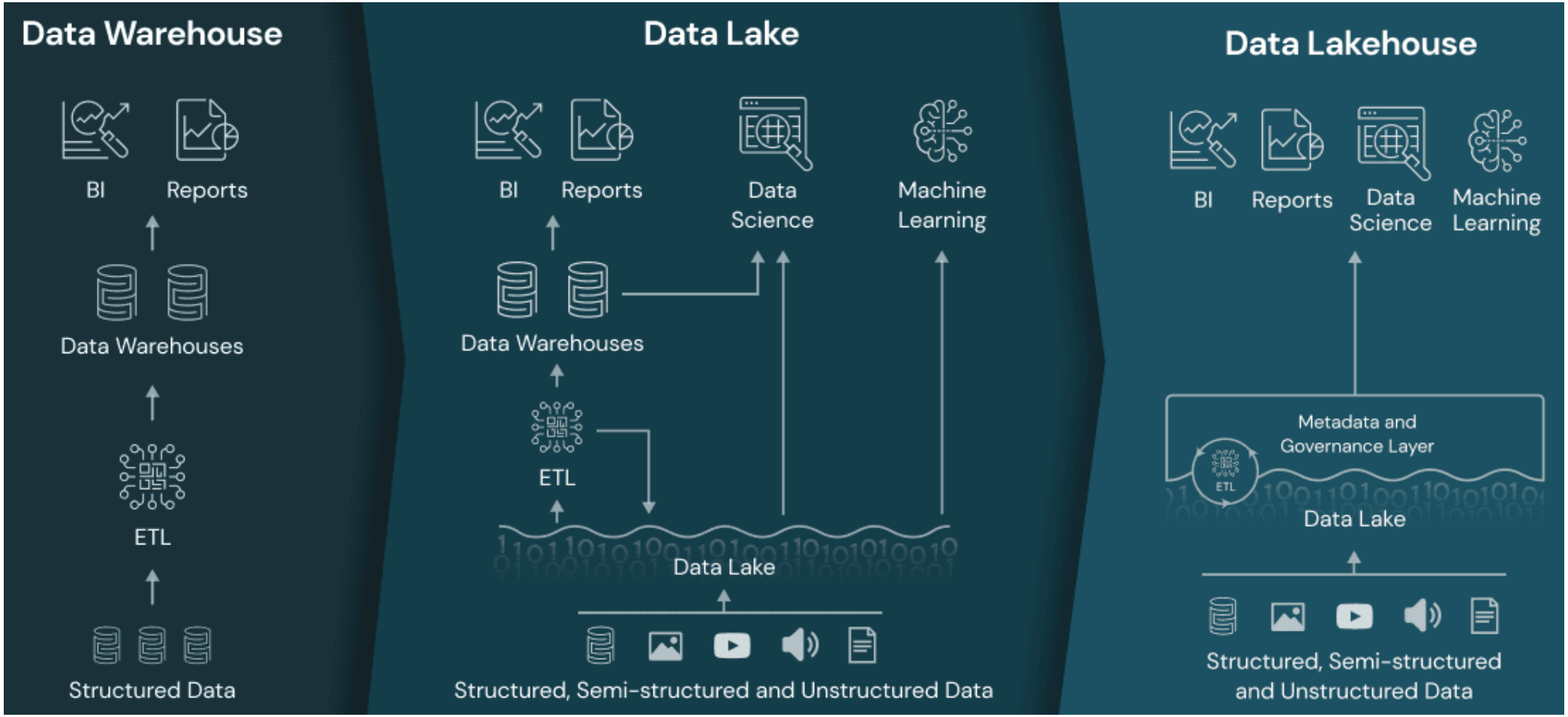
Data lakehouse

	Data warehouse	Data lake	Data lakehouse
Data format	Closed, proprietary format	**Open format** (e.g., Parquet)	Open format
Types of data	Structured data, with limited support for semi-structured data	**All types**: Structured data, semi-structured data, textual data, unstructured (raw) data	All types: Structured data, semi-structured data, textual data, unstructured (raw) data
Data access	SQL-only, no direct access to file	**Open APIs** for direct access to files with SQL, R, Python and other languages	Open APIs for direct access to files with SQL, R, Python and other languages
Reliability	**High quality**, reliable data with ACID transactions	Low quality, data swamp	High quality, reliable data with ACID transactions
Governance and security	**Fine-grained** security and governance for row/columnar level for tables	Poor governance as security needs to be applied to files	Fine-grained security and governance for row/columnar level for tables
Performance	**High**	Low	High
Scalability	Scaling becomes exponentially more expensive	**Scales** to hold any amount of data at low cost, regardless of type	Scales to hold any amount of data at low cost, regardless of type
Use case support	Limited to BI, SQL applications and decision support	Limited to machine learning	One data architecture for BI, SQL and machine learning

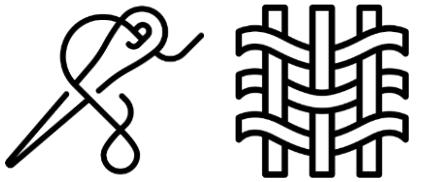


https://databricks.com/blog/2021/05/19/evolution-to-the-data-lakehouse.html

Data lakehouse



Data fabric



Data fabric enables frictionless access and sharing of data in a distributed data environment

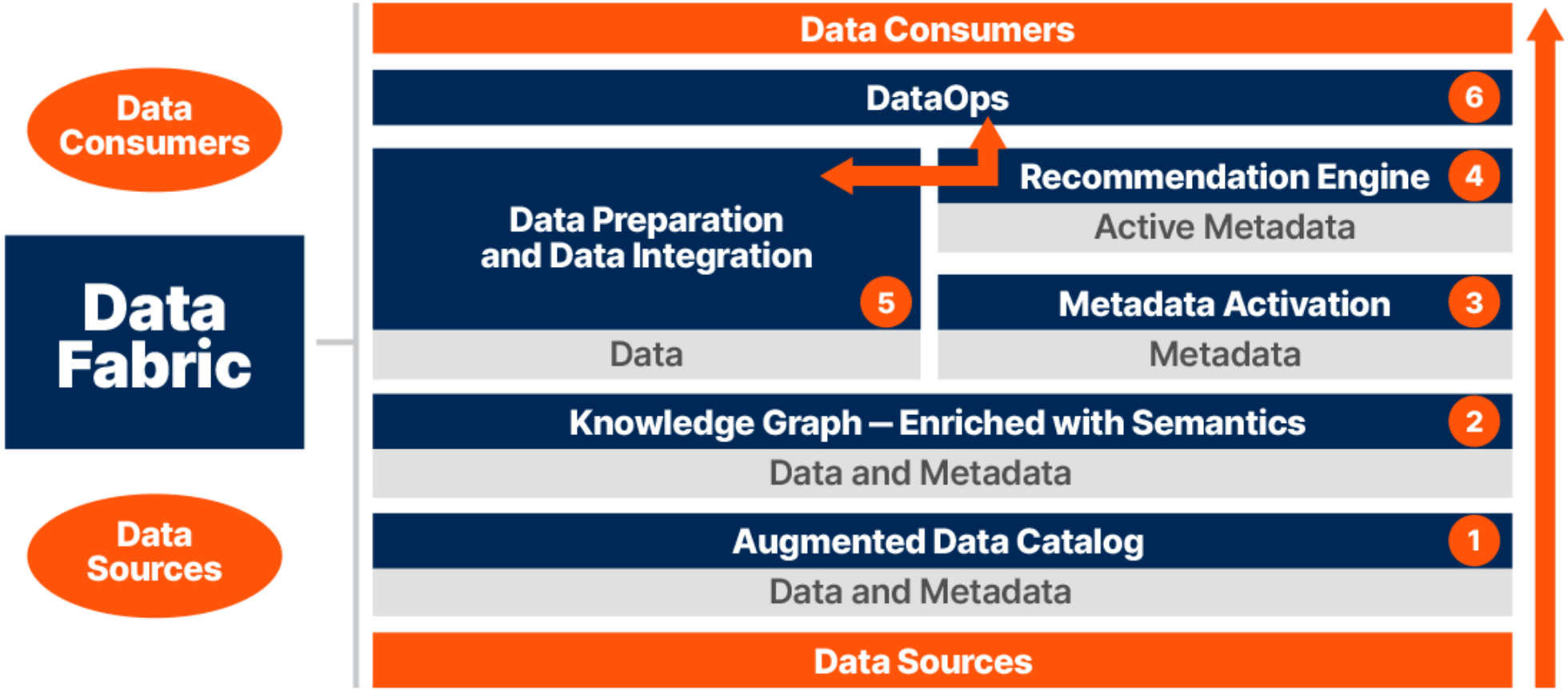
- It enables **a single and consistent data management framework**, which allows seamless data access and processing by design across otherwise siloed storage
- Leverages **both human and machine capabilities** to access data in place or support its consolidation where appropriate
- **Continuously** identifies and connects data from disparate applications to discover unique, business-relevant relationships between the available data points

It is a unified architecture with an integrated set of technologies and services

- Designed to deliver **integrated and enriched data** – at the right time, in the right method, and to the right data consumer – in support of both operational and analytical workloads
- Combines key data management technologies – such as data catalog, data governance, data integration, data pipelining, and data orchestration

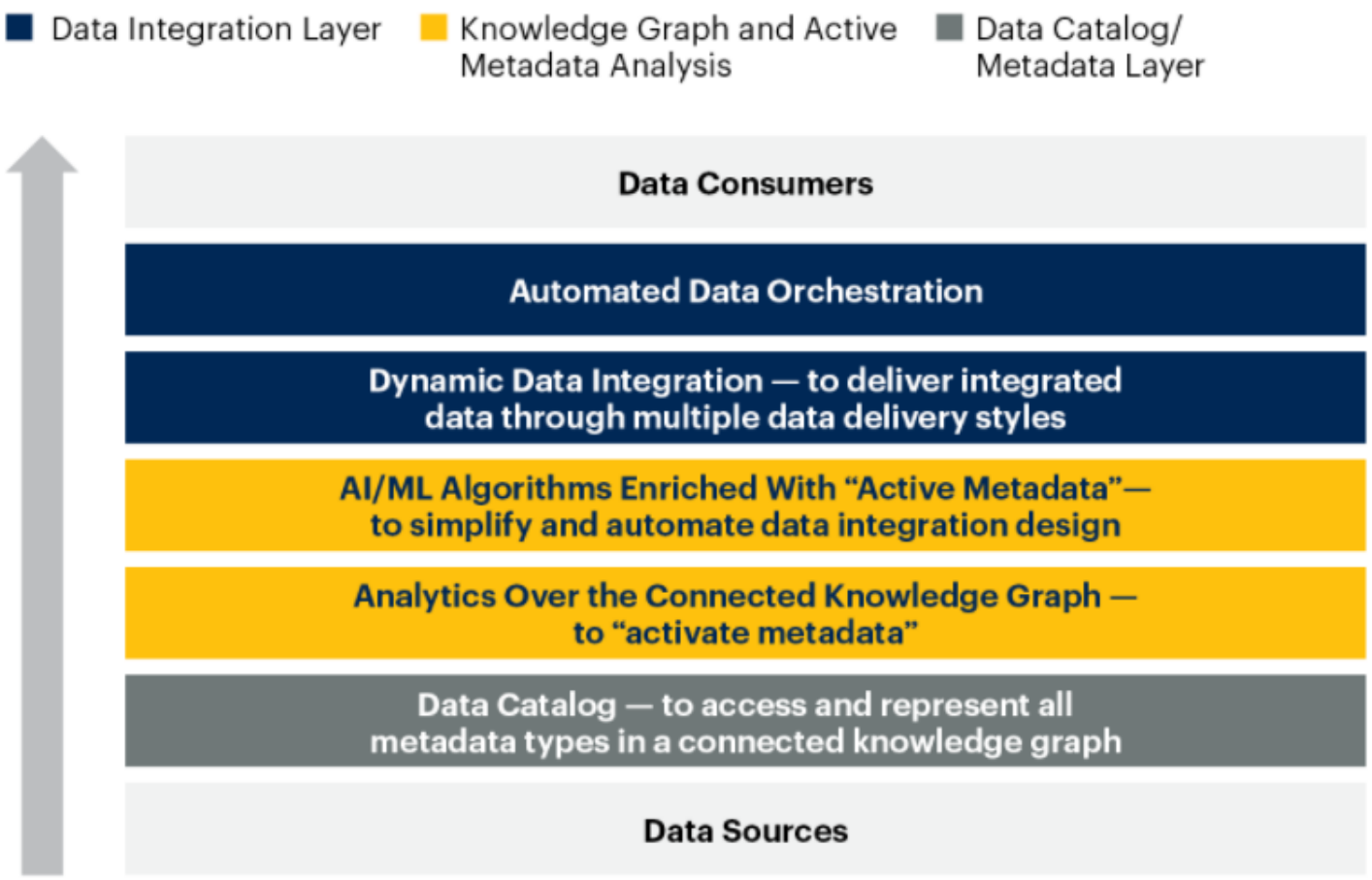
Gartner, 2019 https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo
Gartner, 2021 https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration
K2View Whitepaper: What is a Data Fabric? The Complete Guide, 2021

Data fabric



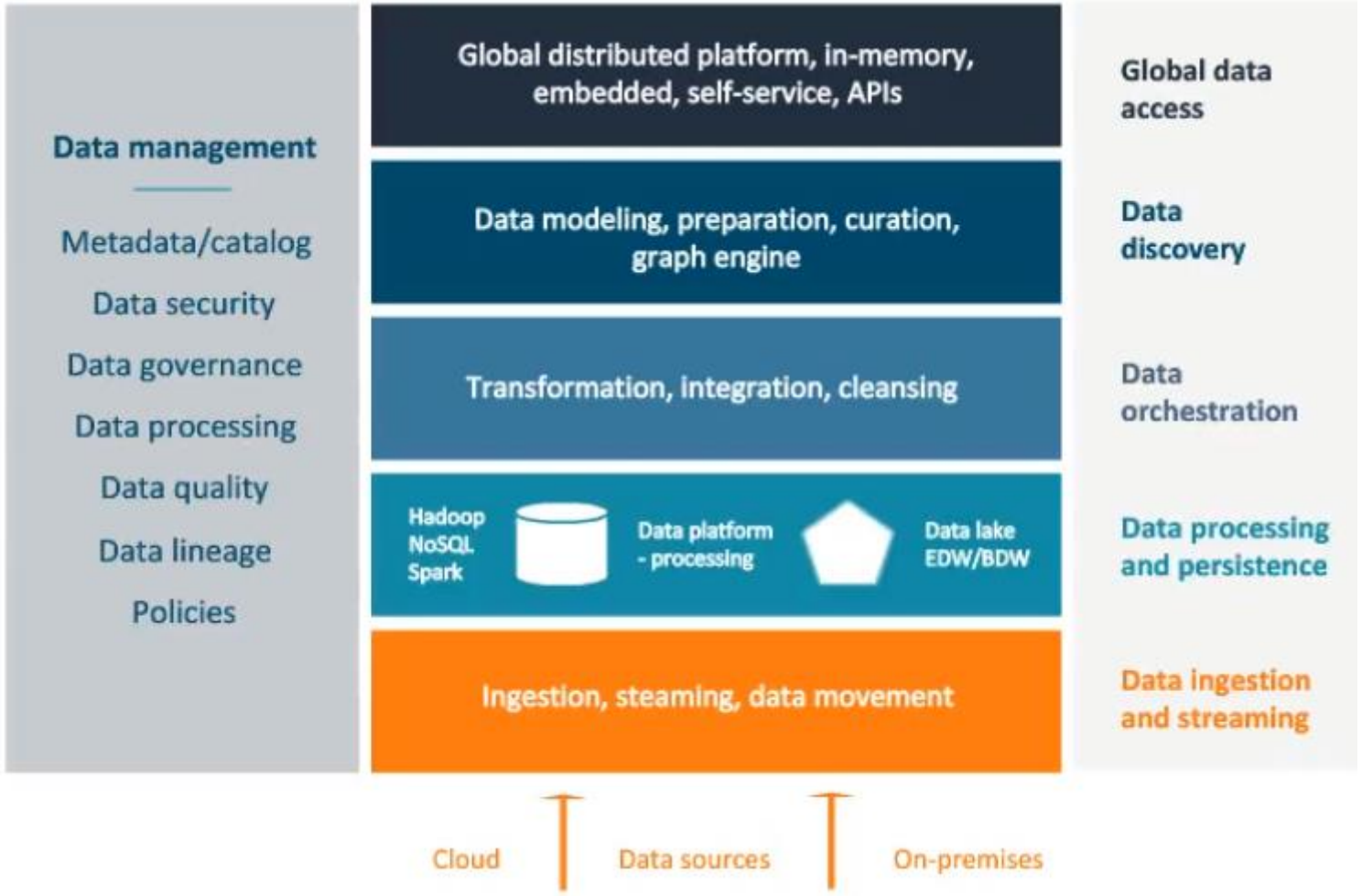
https://www.irion-edm.com/data-management-insights/gartner-data-summit-irion-representative-vendor-for-data-fabric-technology/

Data fabric



Gartner, 2021 https://www.gartner.com/smarterwithgartner/data-fabric-architecture-is-key-to-modernizing-data-management-and-integration

Data fabric


Data management
Metadata/catalog
Data security
Data governance
Data processing
Data quality
Data lineage
Policies
Global distributed platform, in-memory, embedded, self-service, APIs
Data modeling, preparation, curation, graph engine
Transformation, integration, cleansing
Hadoop NoSQL Spark
Data platform - processing
Data lake EDW/BDW
Ingestion, steaming, data movement
Global data access
Data discovery
Data orchestration
Data processing and persistence
Data ingestion and streaming
Cloud
Data sources
On-premises