BIG DATA AND CLOUD PLATFORMS

From databases to data platforms

How did we get here?

Data-Driven Innovation

- Use of data and analytics to foster new products, processes and markets
- Drive discovery and execution of innovation, achieving new services with a business value

Analytics

- A catch-all term for different business intelligence (BI)- and application-related initiatives
  - E.g., of analyzing information from a particular domain
  - E.g., applying BI capabilities to a specific content area (e.g., sales, service, supply chain)

Advanced Analytics

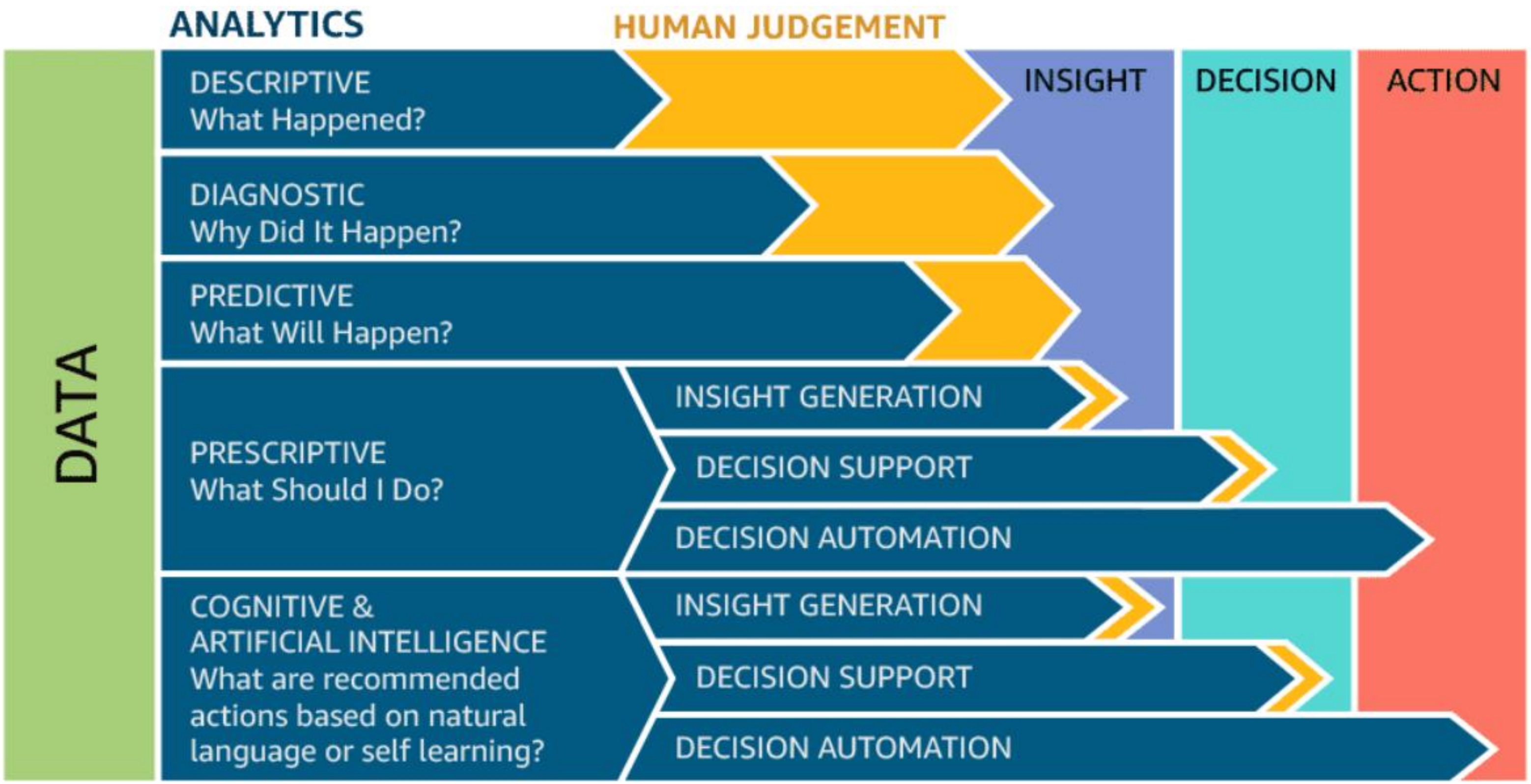
- (Semi-)Autonomous examination of data to discover deeper insights, make predictions, or generate recommendations (e.g., through data/text mining and machine learning)

Augmented Analytics

- Use of technologies such as machine learning and AI to assist with data preparation, insight generation and insight explanation to augment how people explore and analyze data

https://www.gartner.com/en/information-technology/glossary (accessed 2022-08-01)

How did we get here?



Data platform

Companies are collecting tons of data to enable advanced analytics

- Raw data is difficult to obtain, interpret, and maintain
- Data is more and more heterogeneous
- There is need for curating data to make it consumable

Where are we collecting/processing data?

- Getting value from data is not (only) a matter of storage
- Need integrated and multilevel analytical skills and techniques

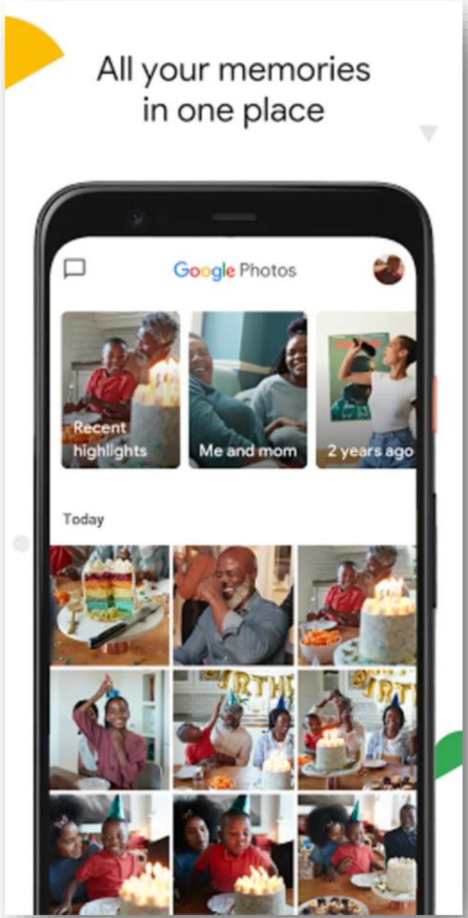
Data platform

> “It is a capital mistake to theorize before one has data. Insensibly, one begins to twist the facts to suit theories, instead of theories to suit facts.”
>
> – Sherlock Holmes

Getting value from data is not (only) a matter of storage

- Any example?

Case study: photo gallery



Search by people, things & places in your photos

Search your photos for anything. For example, you can search for:

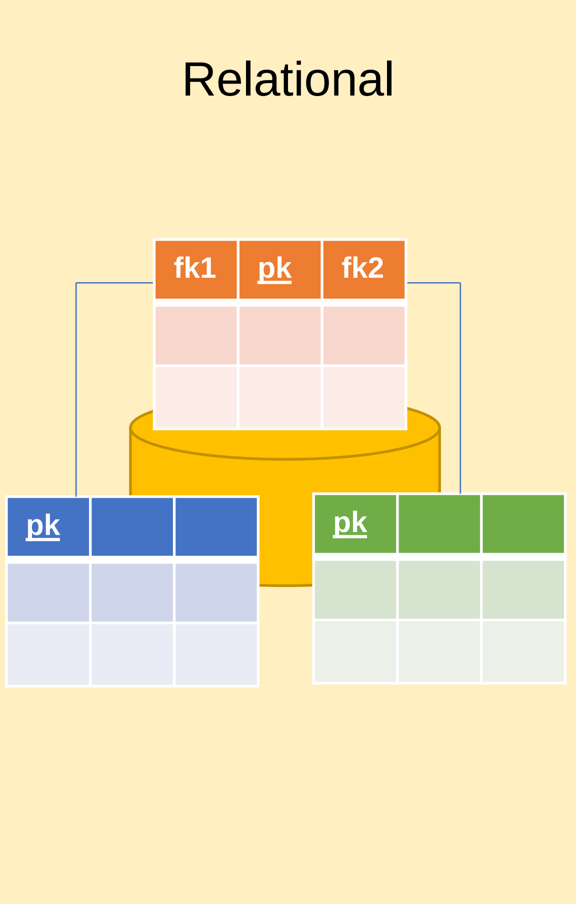
- A wedding you attended last summer
- Your best friend
- A pet
- Your favorite city

Important: Some features are not available in all countries, all domains, or all account types.

Data platform

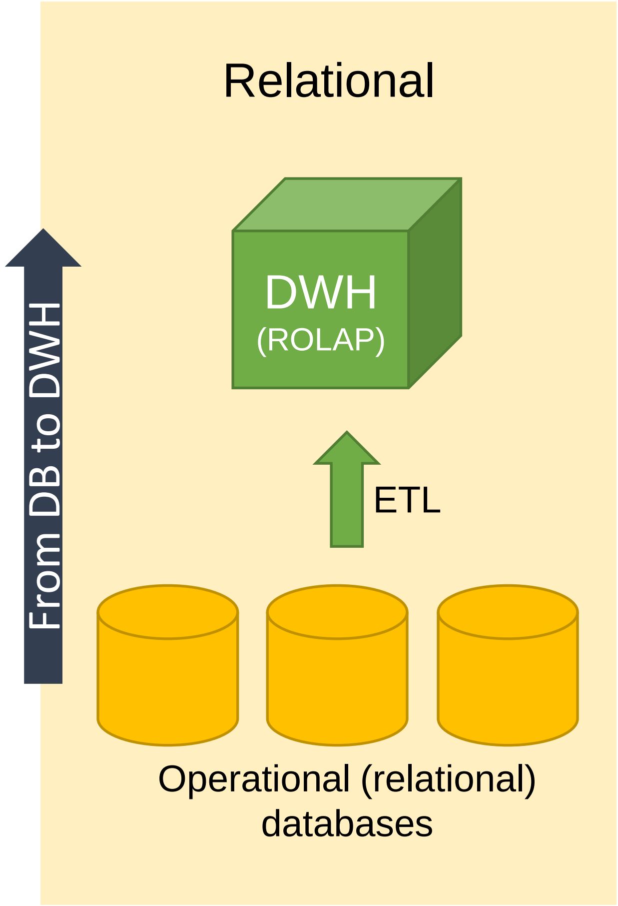
Database

"A database is a structured and persistent collection of information about some aspect of the real world organized and stored in a way that facilitates efficient retrieval and modification. The structure of a database is determined by an abstract data model. Primarily, it is this structure that differentiates a database from a data file."



Özsu M.T. (2018) Database. In: Encyclopedia of Database Systems. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-8265-9_80734

Data platform



Data Warehouse

"A collection of data that supports decision-making processes. It provides the following features: subject-oriented, integrated and consistent, not volatile."

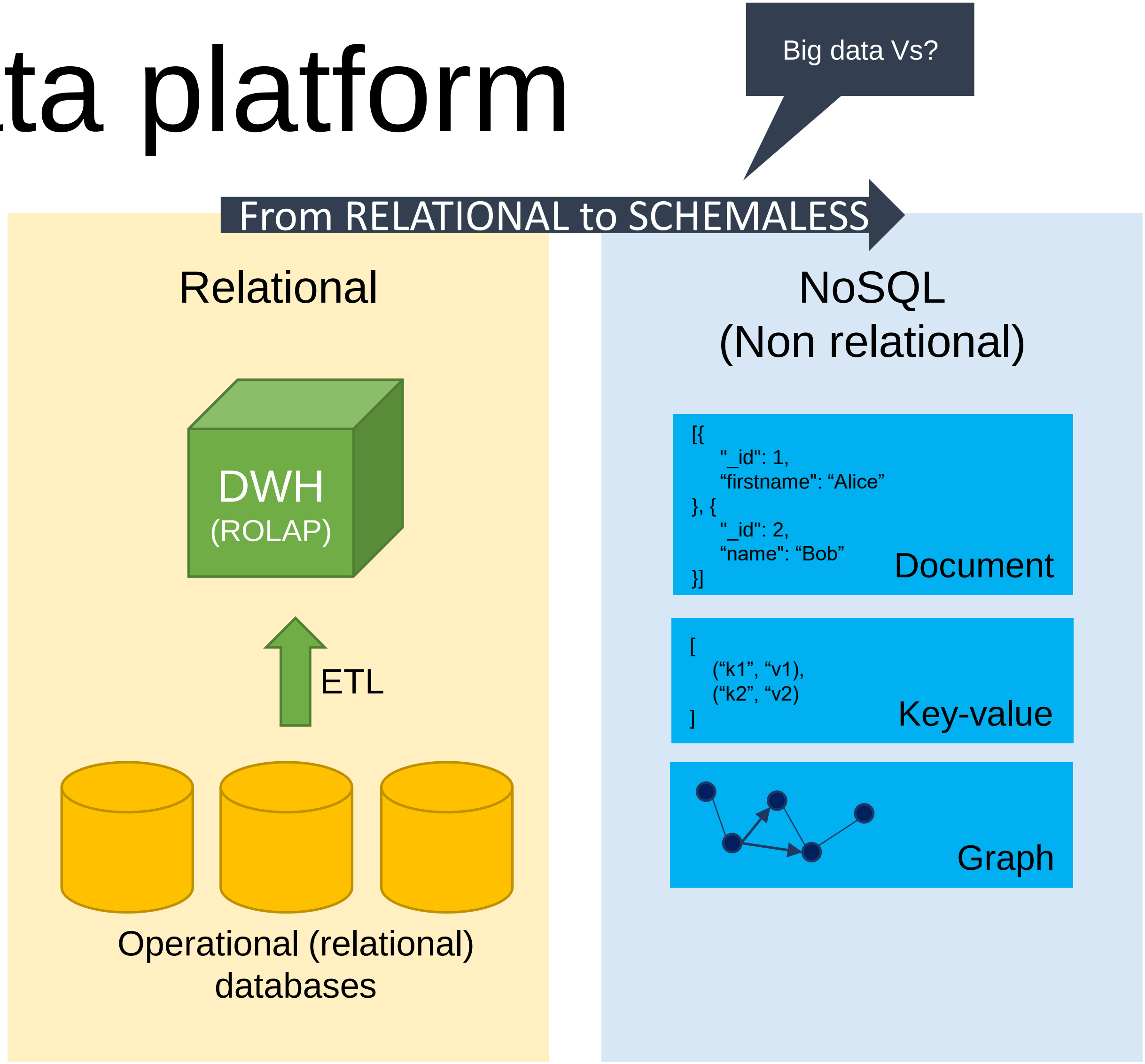
Matteo Golfarelli and Stefano Rizzi. *Data warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc., 2009.

Data platform: OLTP vs OLAP

Data platform: OLTP vs OLAP

Characteristic	OLTP	OLAP
Nature	Constant transactions (queries/updates)	Periodic large updates, complex queries
Examples	Accounting database, online retail transactions	Reporting, decision support
Type	Operational data	Consolidated data
Data retention	Short-term (2-6 months)	Long-term (2-5 years)
Storage	Gigabytes (GB)	Terabytes (TB) / Petabytes (PB)
Users	Many	Few
Protection	Robust, constant data protection and fault tolerance	Periodic protection

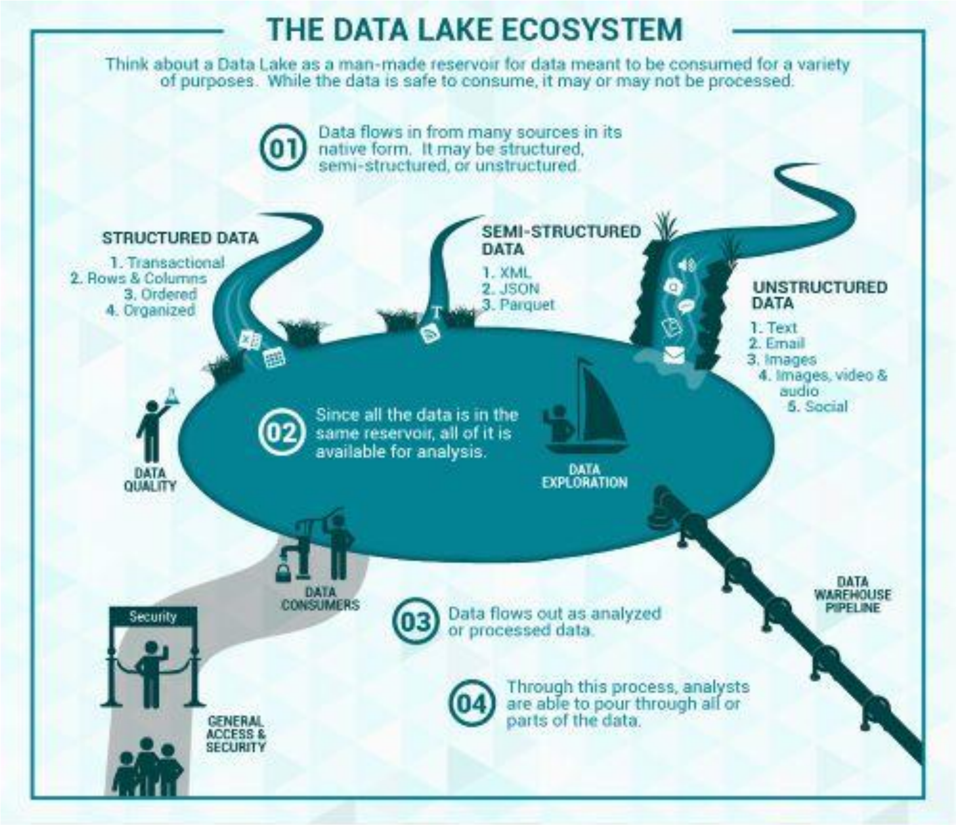
Data platform



Data platform

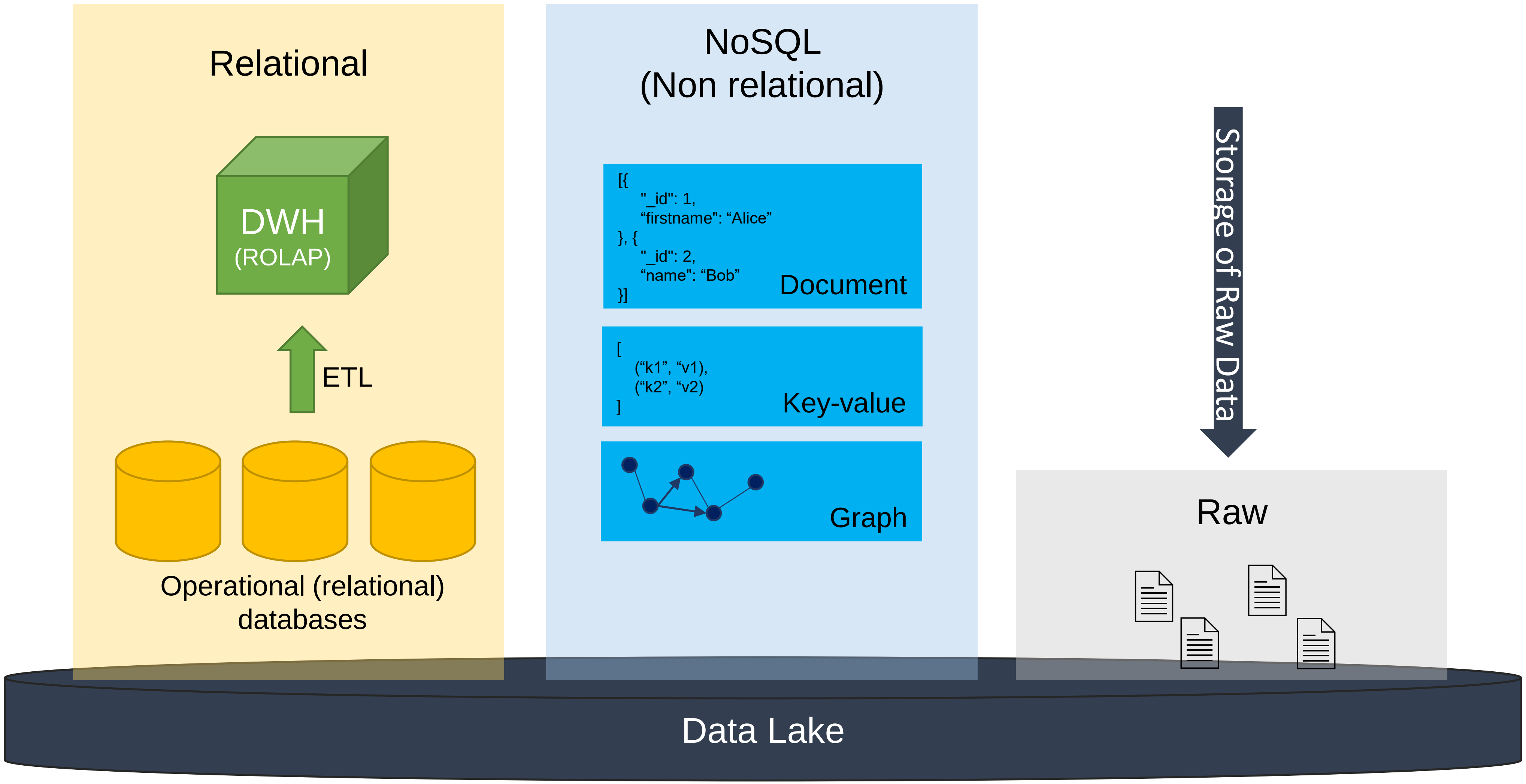
Data lake

Couto et al.: *“A DL is a central repository system for storage, processing, and analysis of raw data, in which the data is kept in its original format and is processed to be queried only when needed. It can store a varied amount of formats in big data ecosystems, from unstructured, semi-structured, to structured data sources”*



Couto, Julia, et al. "A Mapping Study about Data Lakes: An Improved Definition and Possible Architectures." *SEKE*. 2019.
https://dunnsolutions.com/business-analytics/big-data-analytics/data-lake-consulting

Data platform



Data platform: DWH vs Data Lake

Data platform: DWH vs Data Lake

Characteristics	Data warehouse	Data lake
Data	Relational	Non-relational and relational
Schema	Designed prior to implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/ performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data quality	Highly curated data that serves as the central version of the truth	Any data, which may or may not be curated (e.g., raw data)
Users	Business analysts	Data scientists, data developers, and business analysts (using curated data)
Analytics	Batch reporting, BI, and visualizations	Machine learning, predictive analytics, data discovery, and profiling.

Data platform

Data lakes have increasingly taken the role of data hubs

- Eliminate up-front costs of ingestion and ETL since data are stored in original format
- Once in DL, data are available for analysis by everyone in the organization

Drawing a sharp line been storage/computation/analysis is hard

- Is a database just storage?
- What about SQL/OLAP?

Blurring of the architectural borderlines

- DL is often replaced by “data platform” or “data ecosystem”
- Encompass systems supporting data-intensive storage, computation, analysis

Data platform

Data platform

- An integrated set of technologies that collectively meets an organization’s end-to-end data needs such as acquisition, storage, preparation, delivery, and governance, as well as a security layer for users and applications
- Rationale: relieve users from complexity of administration and provision
  - Not only technological skills, but also privacy, access control, etc.
  - Users should only focus on functional aspects

Are we done? No!

- Lacking smart support to govern the complexity of data and transformations
- Data transformations must be governed to prevent DP turning into a swamp
  - Amplified in data science, with data scientists prevailing data architects
  - Leverage descriptive metadata and maintenance to keep control over data

Managing data platforms

Which functionalities for (automated) data management can you think about?

Managing data platforms

- Data provenance
- Compression
- Data profiling
- Entity resolution
- Data versioning
- …

Data provenance

Provenance (also referred to as lineage, pedigree, parentage, genealogy)

- The description of the origins of data and the process by which it arrived at the database
- Not only data products (e.g., tables, files), but also the processes that created them

Examples of use cases

- Business domain. *Users traditionally work with an organized data schema, where the structure and semantics of the data in use is shared across the corporation or even B2B. Yet, a large proportion of businesses deal with bad quality data. Sources of bad data need to be identified and corrected to avoid costly errors in business forecasting.*
- Scientific/research domain. *Data used in the scientific field can be ad hoc and driven by individual researchers or small communities. The scientific field is moving towards more collaborative research and organizational boundaries are disappearing. Sharing data and metadata across organizations is essential, leading to a convergence on common schemes to ensure compatibility. Issues of trust, quality, and copyright of data are significant when using third-party data in such a loosely connected network.*

Simmhan, Yogesh L., Beth Plale, and Dennis Gannon. "A survey of data provenance techniques." *Computer Science Department, Indiana University, Bloomington IN* 47405 (2005): 69.

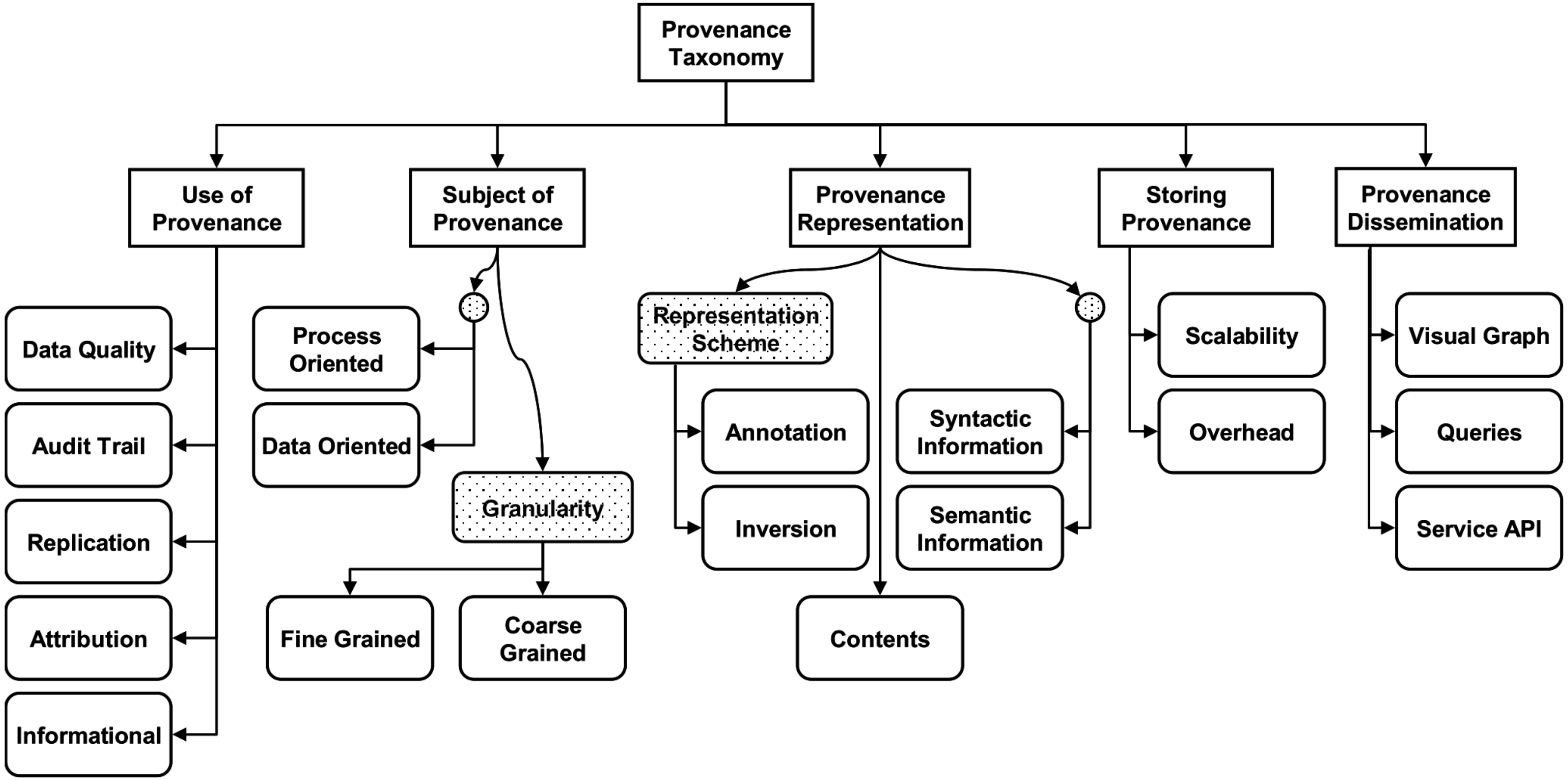
Data provenance

Astronomers are creating an international Virtual Observatory

- A federation of all the world significant astronomical data resources coupled with provision of the computational resources needed to exploit the data scientifically
- Astronomy changed from being an individualistic to a collective enterprise
- Telescope time is devoted/allocated to systematic sky surveys and analysis is performed using data from the archives
- Astronomers are increasingly relying on data that they did not take themselves
- Raw data bear many instrumental signatures that must be removed in the process of generating data products

Mann, Bob. "Some data derivation and provenance issues in astronomy." *Workshop on Data Derivation and Provenance, Chicago*. 2002.
https://www.esa.int/Science_Exploration/Space_Science/Webb/Webb_inspects_the_heart_of_the_Phantom_Galaxy (accessed 2022-08-01)

Data provenance



Simmhan, Yogesh L., Beth Plale, and Dennis Gannon. "A survey of data provenance techniques." *Computer Science Department, Indiana University, Bloomington IN* 47405 (2005): 69.

Data provenance

Granularity

- Fine-grained (instance level): tracking data items (e.g., a tuple in a dataset) transformations
- Coarse-grained (schema-level): tracking dataset transformations

Queries

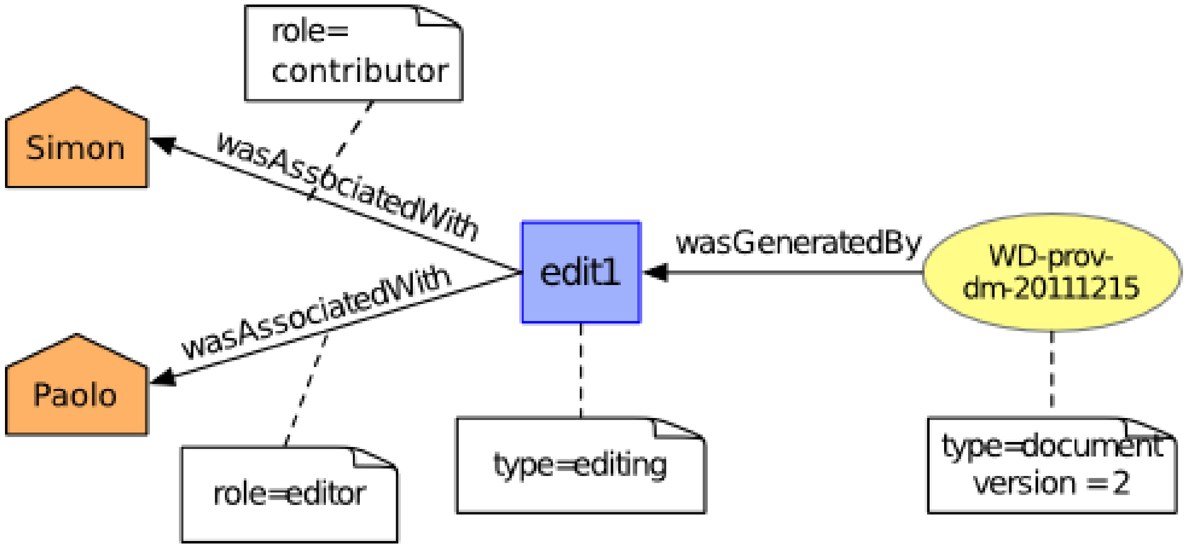
- Where provenance: given some output, which inputs did the output come from?
- How provenance: given some output, how were the inputs manipulated?
- Why provenance: given some output, why was data generated?
  - E.g., in the form of a proof tree that locates source data items contributing to its creation

Simmhan, Yogesh L., Beth Plale, and Dennis Gannon. "A survey of data provenance techniques." *Computer Science Department, Indiana University, Bloomington IN* 47405 (2005): 69.
Ikeda, Robert, and Jennifer Widom. *Data lineage: A survey*. Stanford InfoLab, 2009.

Data provenance

Data provenance, an example of data management

- Metadata pertaining to the history of a data item
- Pipeline including the origin of objects and operations they are subjected to
- We have a standard: https://www.w3.org/TR/prov-dm/



https://www.w3.org/TR/prov-dm/

Data provenance

Entity

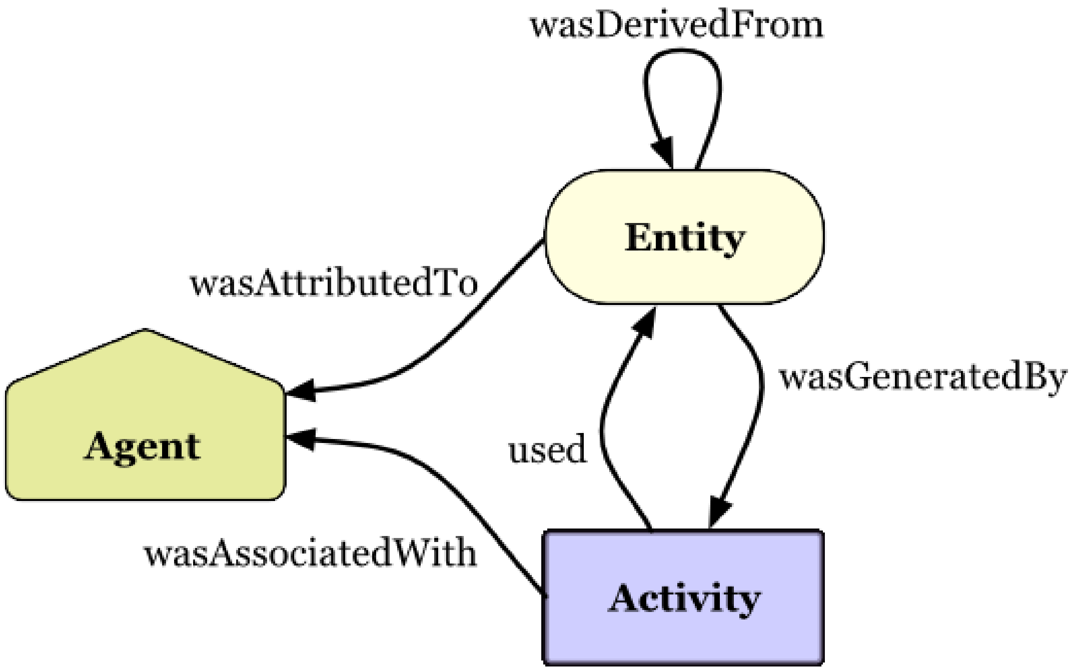
- Physical/conceptual things

Activity

- Dynamic aspects of the world, such as actions
- How entities come into existence, often making use of previously existing entities

Agent

- A person, a piece of software
- Takes a role in an activity such that the agent can be assigned some degree of responsibility for the activity taking place



https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

Data provenance

Use cases for data provenance

Accountability and auditing

Data quality

- Monitoring of the quality (e.g., accuracy) of the objects produced
- Notify when a transformation pipeline is not behaving as expected

Debugging

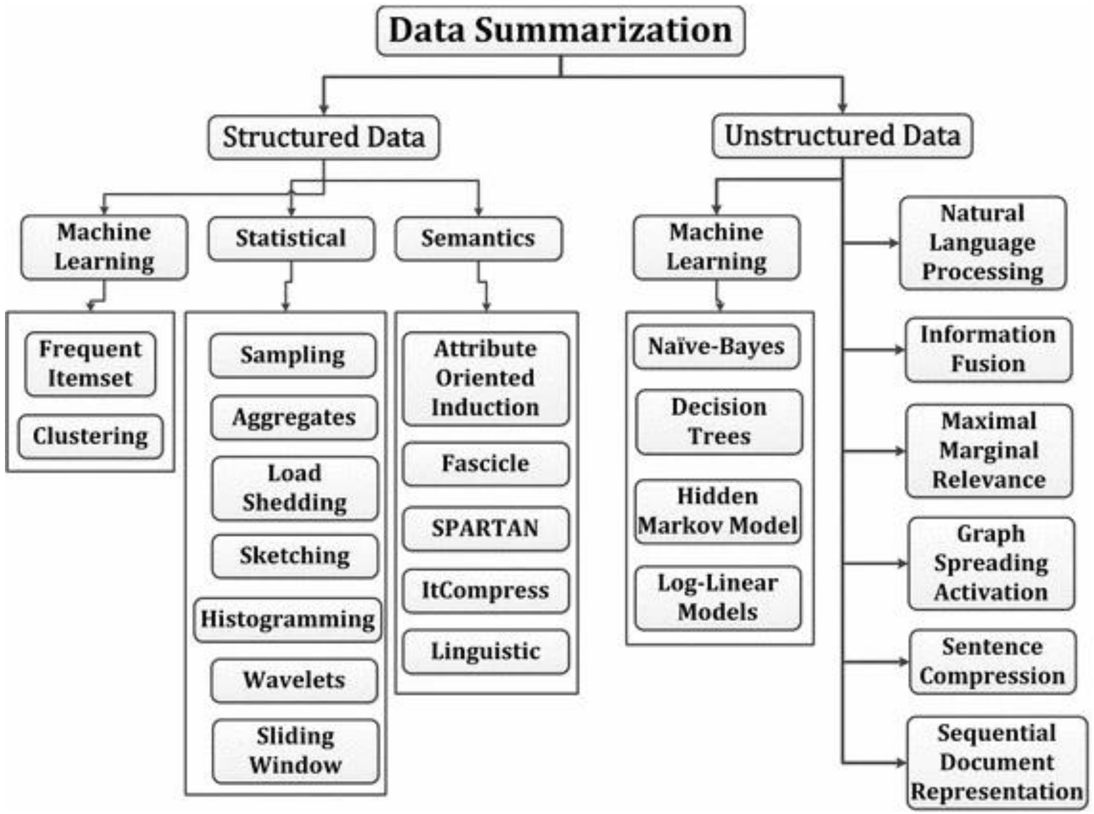
- Inferring the cause of pipeline failures is challenging
- Store inputs of each operation with versions and environmental settings (RAM, CPUs, etc.)

And so on...

Compression

Summarization / compression

- Present a concise representation of a dataset in a comprehensible and informative manner

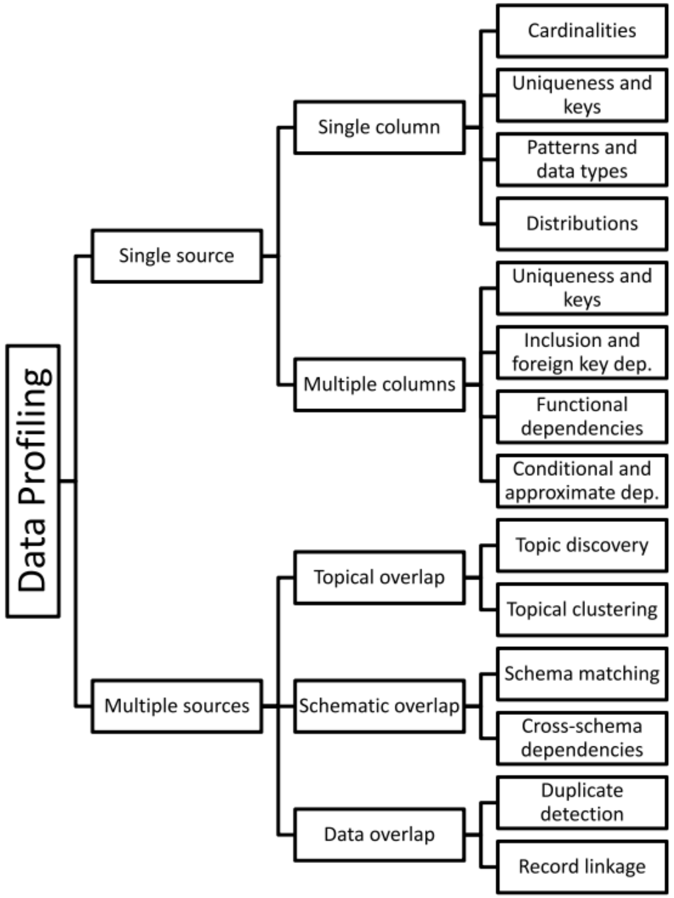


Ahmed, Mohiuddin. "Data summarization: a survey." *Knowledge and Information Systems* 58.2 (2019): 249-273.

Data profiling

Data profiling

- A broad range of methods to efficiently analyze a given data set
- E.g., in a relational scenario, tables of a relational database are scanned to derive metadata, such as data types and value patterns, completeness and uniqueness of columns, keys and foreign keys, and occasionally functional dependencies and association rules



Naumann, Felix. "Data profiling revisited." *ACM SIGMOD Record* 42.4 (2014): 40-49.

Data profiling

Use cases

- Query optimization
  - Performed by DBMS to support query optimization with statistics about tables and columns
  - Profiling results can be used to estimate the selectivity of operators and the cost of a query plan
- Data cleansing (typical use case is profiling data)
  - Prepare a cleansing process by revealing errors (e.g., in formatting), missing values or outliers
- Data integration and analytics

Challenges?

Naumann, Felix. "Data profiling revisited." *ACM SIGMOD Record* 42.4 (2014): 40-49.

Data profiling

a	b	c	d
1	1	2	2
1	2	1	4

Challenges

- The results of data profiling are computationally complex to discover
  - E.g., discovering keys/dependencies usually involves some sorting step for each considered column
- Verification of complex constraints on column combinations in a database
  - What is the complexity of this task?

Complexity

- Given a table with columns $C = \{ a, b, c, d \}$
- To extract the (distinct) cardinality of each column, I will consider |C| columns
  (a), (b), (c), (d)
- To extract the correlations between pairs of columns, I will consider $\binom{|C|}{2}$ groups
  (a, b), (a, c), (a, d), (b, c), (c, d), (c, d)
- Extracting the relationships among all possible groups of columns generalizes to $\sum_{n=1}^{|C|}\binom{|C|}{n} = 2^{|C|} - 1$ groups

Naumann, Felix. "Data profiling revisited." *ACM SIGMOD Record* 42.4 (2014): 40-49.

Entity resolution

Entity resolution

- (also known as entity matching, linking)
- Find records that refer to the same entity across different data sources (e.g., data files, books, websites, and databases)

ID	Name	Telephone	Address	Items Purchased
233	Angelica J. Jordan	334-555-0178	111 Spring Ln, Greenville, AL	5556, 7611
452	Angie Jordan	202-555-5477	45 Krakow St, Washington, DC	2297
699	Andrew Jordan	334-555-0178	111 Spring Ln, Greenville, AL	1185, 2299, 3720
720	Angie Jrodon			5556
821	Angelica Jeffries Jordan	202-555-5477	397 Hope Blvd, Greenville, AL	7611

Papadakis, George, et al. "Blocking and filtering techniques for entity resolution: A survey." *ACM Computing Surveys (CSUR)* 53.2 (2020): 1-42.

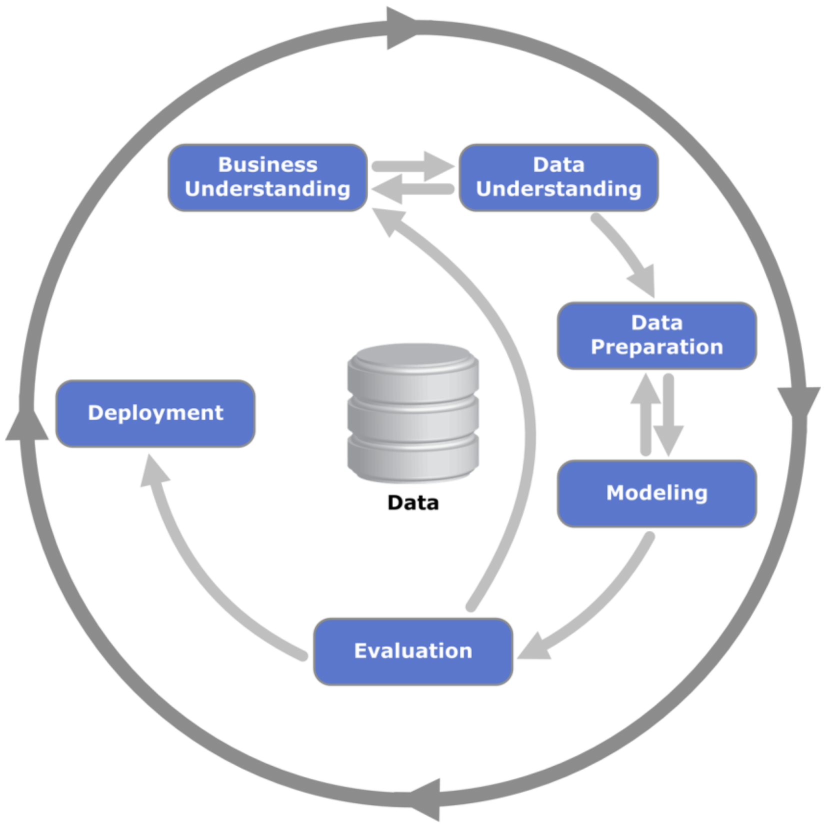
Data versioning

Version control

- A class of systems responsible for managing changes to computer programs, documents, or data collections
- Changes are identified by a number/letter code, termed the revision/version number

However, data pipelines are not only about code bult also about

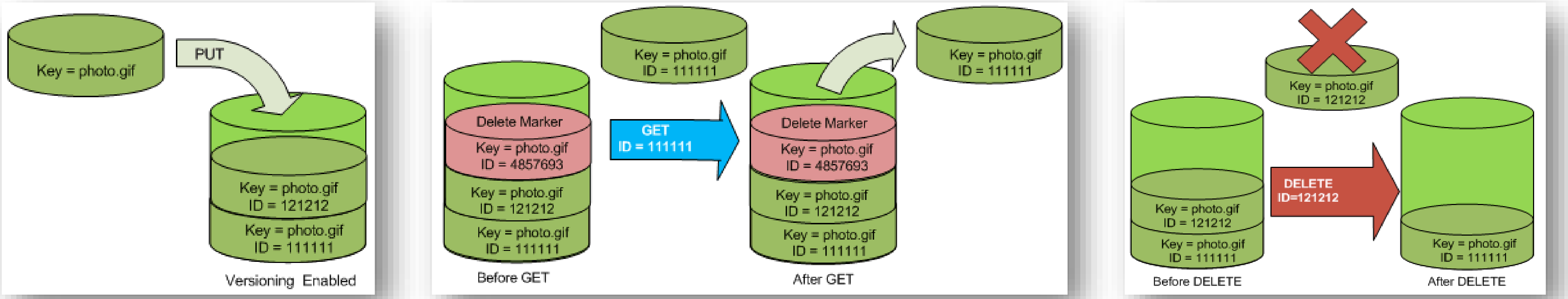
- Model Version control
- Data Version Control
- Model Parameter Tracking
- Model Performance Comparison



Data versioning

Support CRUD (Create, Read, Update, Delete) operations with versions

E.g., on AWS (PUT, GET, DELETE), what about update?



https://docs.aws.amazon.com/AmazonS3/latest/userguide/versioning-workflows.html (accessed 2022-08-01)

Data platform

Are we done? No!

- Metadata can become bigger than data themselves

We need meta meta-data (or models)...

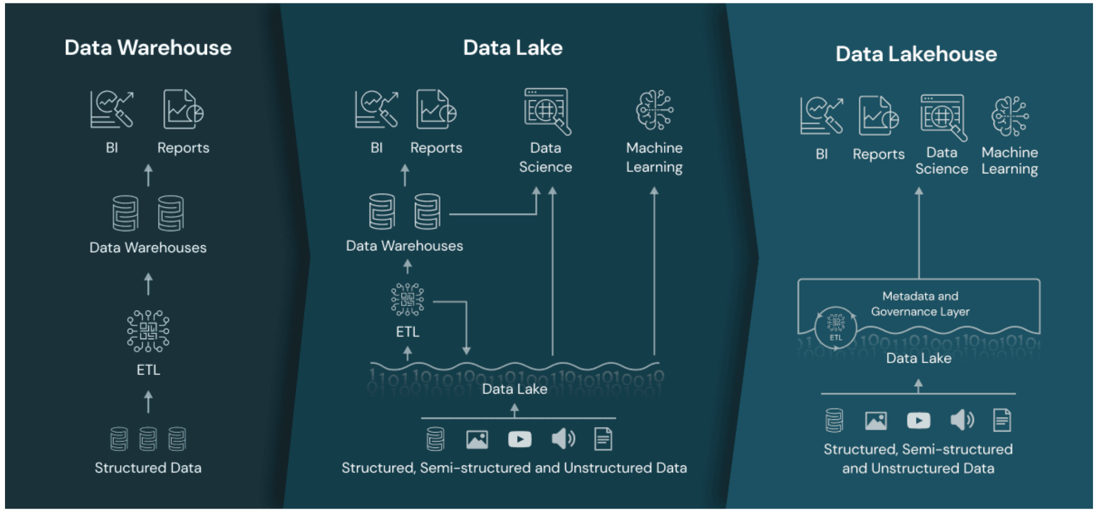
- ... chasing our own tails

Data management is still a (research) issue in data platforms

Data lakehouse

Data lakehouse

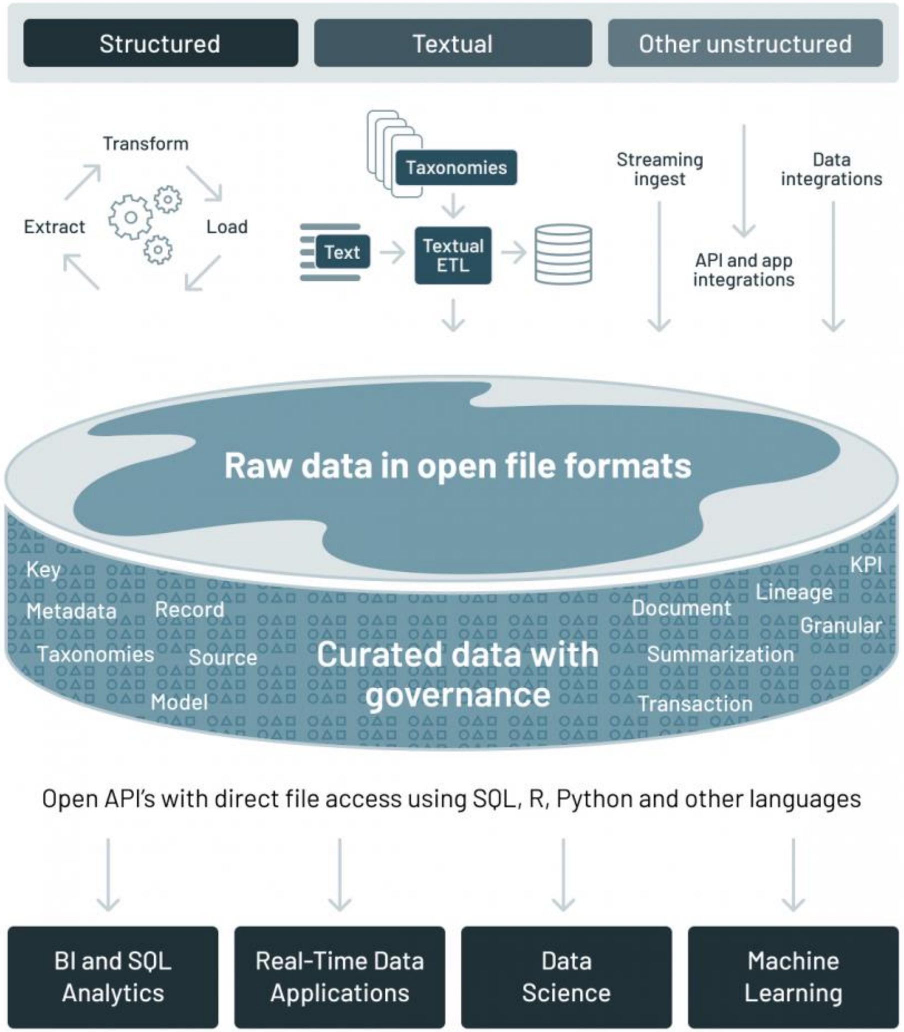
- Data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data
- Vendor lock in



https://www.databricks.com/glossary/data-lakehouse

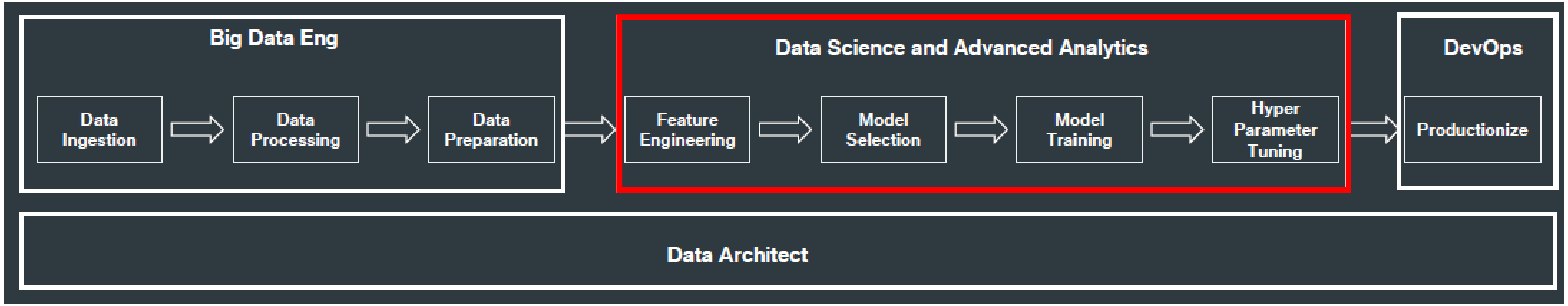
Data lakehouse

	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 platform



Data platform: related job positions

Data platform engineer

- Orchestrate the successful implementation of cloud technologies within the data infrastructure of their business
- Solid understanding of impact database types and implementation
- Responsible for purchasing decisions for cloud services and approval of data architectures

Data architect

- Team members who understand all aspects of a data platform's architecture
- Work closely with the data platform engineers to create data workflows
- Responsible for designing and testing new database architectures and planning both data and architecture migrations

Data pipeline engineer

- Responsible for planning, architecting, and building large-scale data processing systems

Data analyst

- Analyze data systems, creating automated systems for retrieving data from the data platform
- Cloud data analysts are more commonly members of the business user population

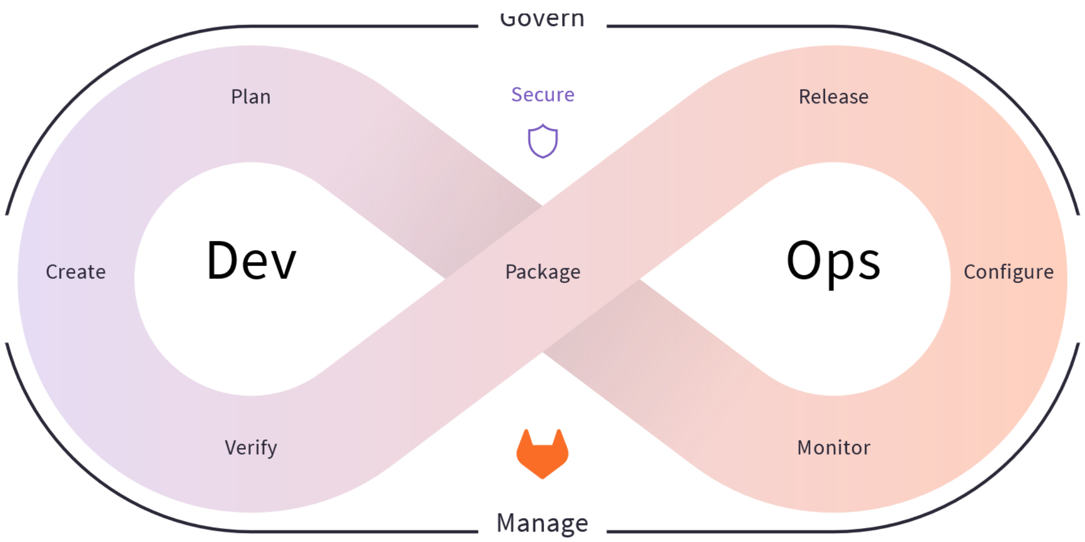
Data scientist

- Analyze and interpret complex digital data
- Work with new technologies (e.g., machine learning) to deepen the business' understanding and gain new insights

From DevOps…

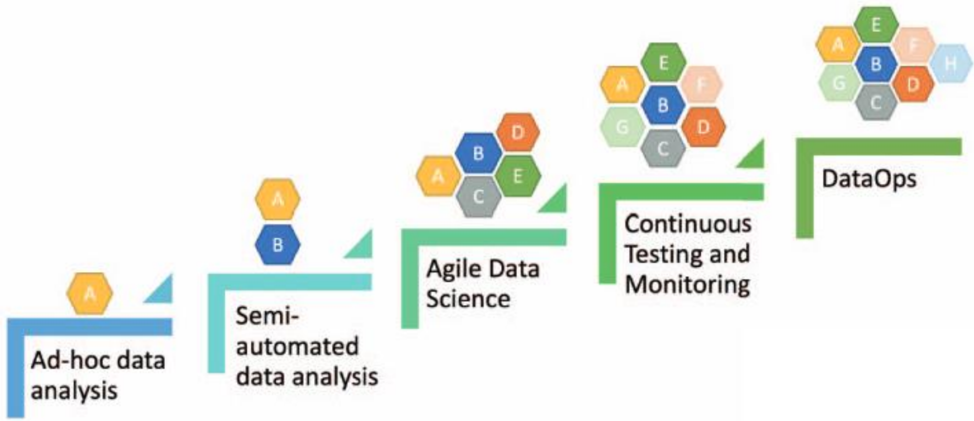
DevOps combines development and operations to increase the efficiency, speed, and security of software development and delivery compared to traditional processes.

DevOps practices enable software development (dev) and operations (ops) teams to accelerate delivery through automation, collaboration, fast feedback, and iterative improvement



https://about.gitlab.com/topics/devops/ (accessed 2023-06-03)

… to DataOps



Case	Use cases at Ericsson	Interviewed Experts	
		ID	Role
A	Automated data collection for data analytics	R4	Senior Data Scientist
B	Building data pipelines	R1	Integration and Operations Professional
C	Toolkit for Network Analytics	R2	Analytics System Architect
D	Building CI pipelines for Data Scientist team	R7	Data Scientist
E	Tracking the Software Version	R5	Senior Customer Support Engineer
F	Testing the Software Quality	R6	Developer Customer Support
G	KPI Analysis Software	R3	Senior Data Engineer
H	Building data pipelines for CI and CD data	R8	Program Manager

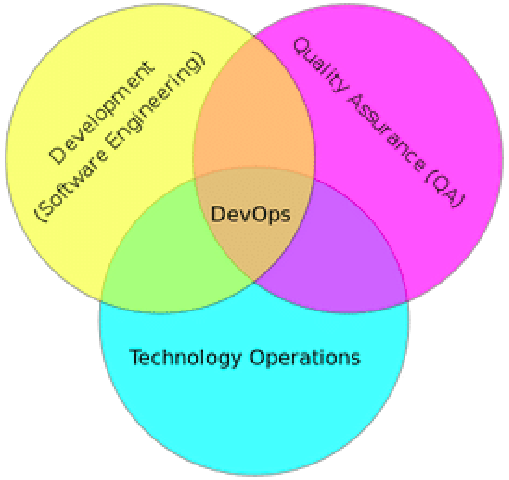
DataOps refers to a general process aimed to shorten the end-to-end data analytic life-cycle time by introducing automation in the data collection, validation, and verification process

Munappy, A. R., Mattos, D. I., Bosch, J., Olsson, H. H., & Dakkak, A. (2020, June). From ad-hoc data analytics to dataops. In *Proceedings of the International Conference on Software and System Processes* (pp. 165-174).

DataOps

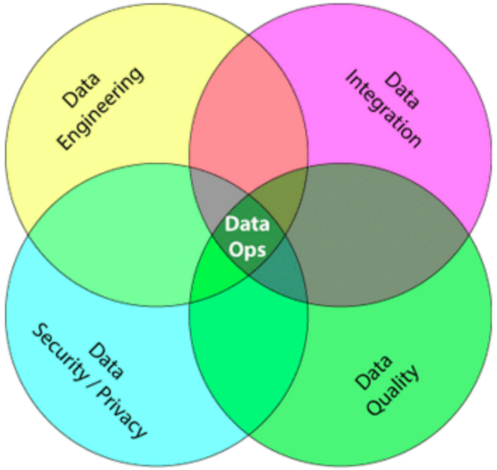
From DevOps to DataOps

- *“A collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization”*
- Data analytics improved in terms of velocity, quality, predictability and scale of software engineering and deployment



Some key rules

- Establish progress and performance measurements at every stage
- Automate as many stages of the data flow as possible
- Establish governance discipline (*governance-as-code*)
- Design process for growth and extensibility



Gartner, 2020 https://www.gartner.com/smarterwithgartner/how-dataops-amplifies-data-and-analytics-business-value
Andy Palmer, 2015 https://www.tamr.com/blog/from-devops-to-dataops-by-andy-palmer/
William Vorhies, 2017 https://www.datasciencecentral.com/profiles/blogs/dataops-it-s-a-secret

Data fabric

“vision for data management […] that seamlessly connects different clouds, whether they are private, public, or hybrid environments.” (2016)

Frictionless access and sharing of data in a distributed data environment

- Enables a **single and consistent data management framework**, which allows seamless data access and processing by design across otherwise siloed storage
- Leverages **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**

https://cloud.netapp.com/hubfs/Data-Fabric/Data%20Fabric%20WP%20April%202017.pdf (accessed 2023-06-23)
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

- **Catalog all your data**: including busi glossa
- **Enable self-service capabilities**: data di
consumption of data-as-a-product
- **Provide a knowledge graph**: Visualizing how da
interconnected, deriving additional actionable ins
- **Provide intelligent (smart) information integra** s
alike in their data integration and transformation,
- **Derive insight from metadata**: Orchestrating ar
integration, data engineering, and data governance end to end
- **Enforce local and global data rules/policies**: Including AI/ML-based automated generation, adjustments, and enforcement of rules and policies
- **Manage an end-to-end unified lifecycle**: Implementing a coherent and consistent lifecycle end to end of all Data Fabric tasks across various platforms, personas, and organizations
- **Enforce data and AI governance**: Broadening the scope of traditional data governance to include AI artefacts, for example, AI models, pipelines

Is this brand new?

Data fabric

It is a design concept

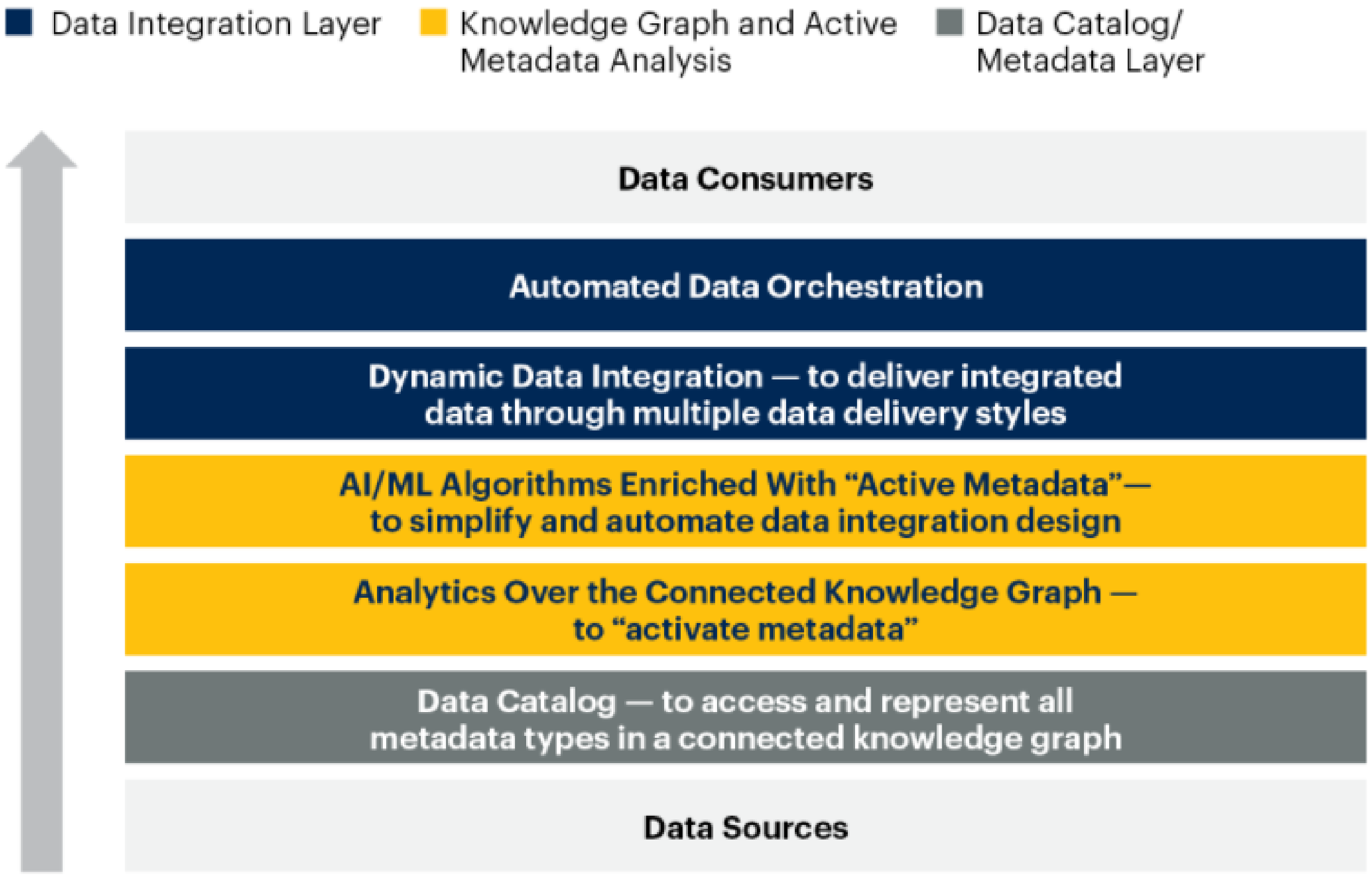
- It optimizes data management by automating repetitive tasks
- According to Gartner estimates, 25% of data management vendors will provide a complete framework for data fabric by 2024 – up from 5% today

Cambridge Semantics	Anzo, AnzoGraph
Cloudera	Cloudera Data Platform
DataRobot	Paxata
Denodo Technologies	Denodo Platform
Hitachi Vantara	Lumada Data Services
IBM	IBM Cloud Pak for Data
Informatica	Informatica Intelligent Data Management
Infoworks	DataFoundry
Oracle	Oracle GoldenGate, Oracle Autonomous Data Platform, Oracle Cloud Infrastructure, Oracle Analytics Cloud
Qlik	Qlik Data Catalyst, Qlik Replicate, Qlik Compose for Data Warehouse, Qlik Compose for Data Lakes
SAP	SAP HANA, SAP Data Intelligence, SAP Information Management, SAP PowerDesigner, SAP Cloud Platform Integration
Solix Technologies	Solix Common Data Platform
Syncsort	Syncsort Connect, Syncsort Trillium, Syncsort Spectrum, Syncsort Ironstream
Talend	Talend Data Fabric
TIBCO Software	TIBCO Unify

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

K2View, 2021 https://www.k2view.com/top-data-fabric-vendors

Data fabric



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

Data mesh

Distributed data architecture, under centralized governance and standardization for interoperability, enabled by a shared and harmonized self-serve data infrastructure

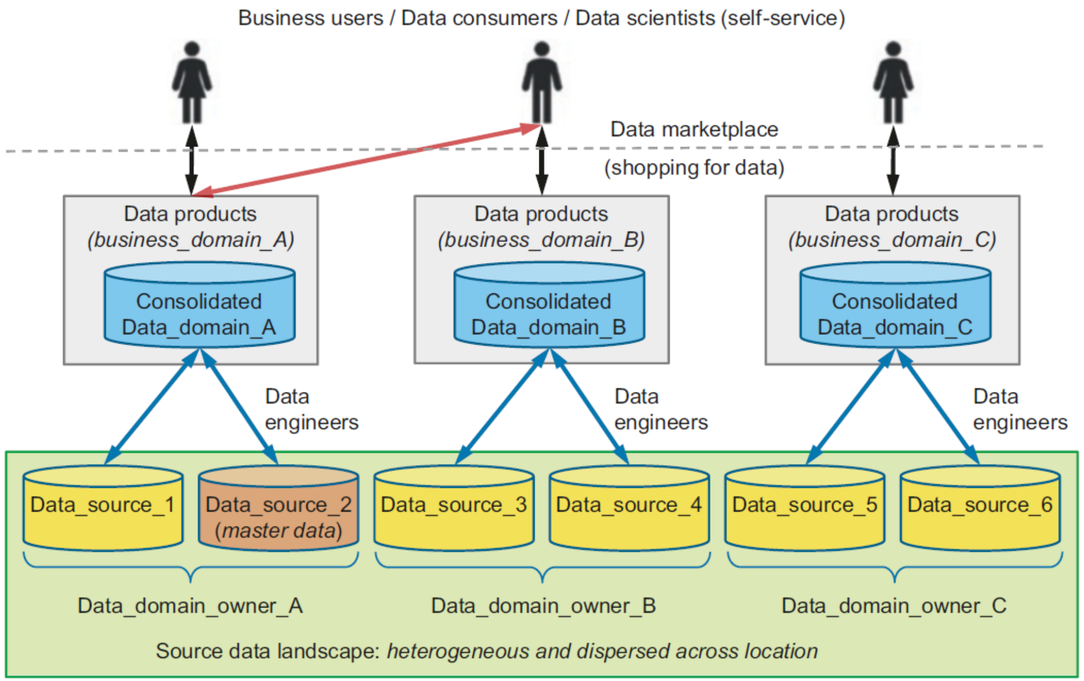
- Domain-oriented decentralized data ownership
  - Decentralization and distribution of responsibility to people who are closest to the data, in order to support continuous change and scalability
  - Each domain exposes its own op/analytical APIs
- **Data as a product** (*quantum*)
  - Products must be discoverable, addressable, trustworthy, self-describing, secure
- Self-serve data infrastructure as a platform
  - High-level abstraction of infrastructure to provision and manage the lifecycle of data products
- Federated computational governance
  - A governance model that embraces decentralization and domain self-sovereignty, interoperability through global standardization, a dynamic topology, automated execution of decisions by the platform

Zhamak Dehghani, 2019 https://martinfowler.com/articles/data-monolith-to-mesh.html
Zhamak Dehghani, 2020 https://martinfowler.com/articles/data-mesh-principles.html

Data mesh

Data Mesh organizes data around **business domain owners** and transforms relevant data assets (data sources) to **data products** that can be consumed by distributed business users from various business domains or functions

- Data products are created, governed, and used in an **autonomous, decentralized**, and self-service manner
- **Self-service capabilities**, which we have already referenced as a Data Fabric capability, enable business organizations to entertain a data marketplace with shopping-for-data characteristics



What makes data a product?

A **data product** is raw data transformed into a business context

- Data products are registered in **knowledge catalog** through specifications (XML, JSON, etc.)
- Main features
  - **Data product description**: The data product needs to be well described
  - **Access methods**: for example, REST APIs, SQL, NoSQL, etc., and where to find the data asset
  - **Policies and rules**: who is allowed to consume the data product for what purpose
  - **SLAs**: agreements regarding the data product availability, performance characteristics, functions, cost of data product usage
  - **Defined format**: A data product needs to be described using a defined format
  - **Cataloged**: All data products need to be registered in the knowledge catalog. Data products need to be searchable and discoverable by potential data product consumers and business user
- Data products themselves are not stored in the knowledge catalog

Data mesh vs data fabric

They are design concepts, not things

- They are not mutually exclusive
- They are architectural frameworks, not architectures
  - The frameworks must be adapted and customized to your needs, data, processes, and terminology
  - Gartner estimates 25% of data management vendors will provide a complete data fabric solution by 2024 – up from 5% today

Alex Woodie, 2021 https://www.datanami.com/2021/10/25/data-mesh-vs-data-fabric-understanding-the-differences/
Dave Wells, 2021 https://www.eckerson.com/articles/data-architecture-complex-vs-complicated

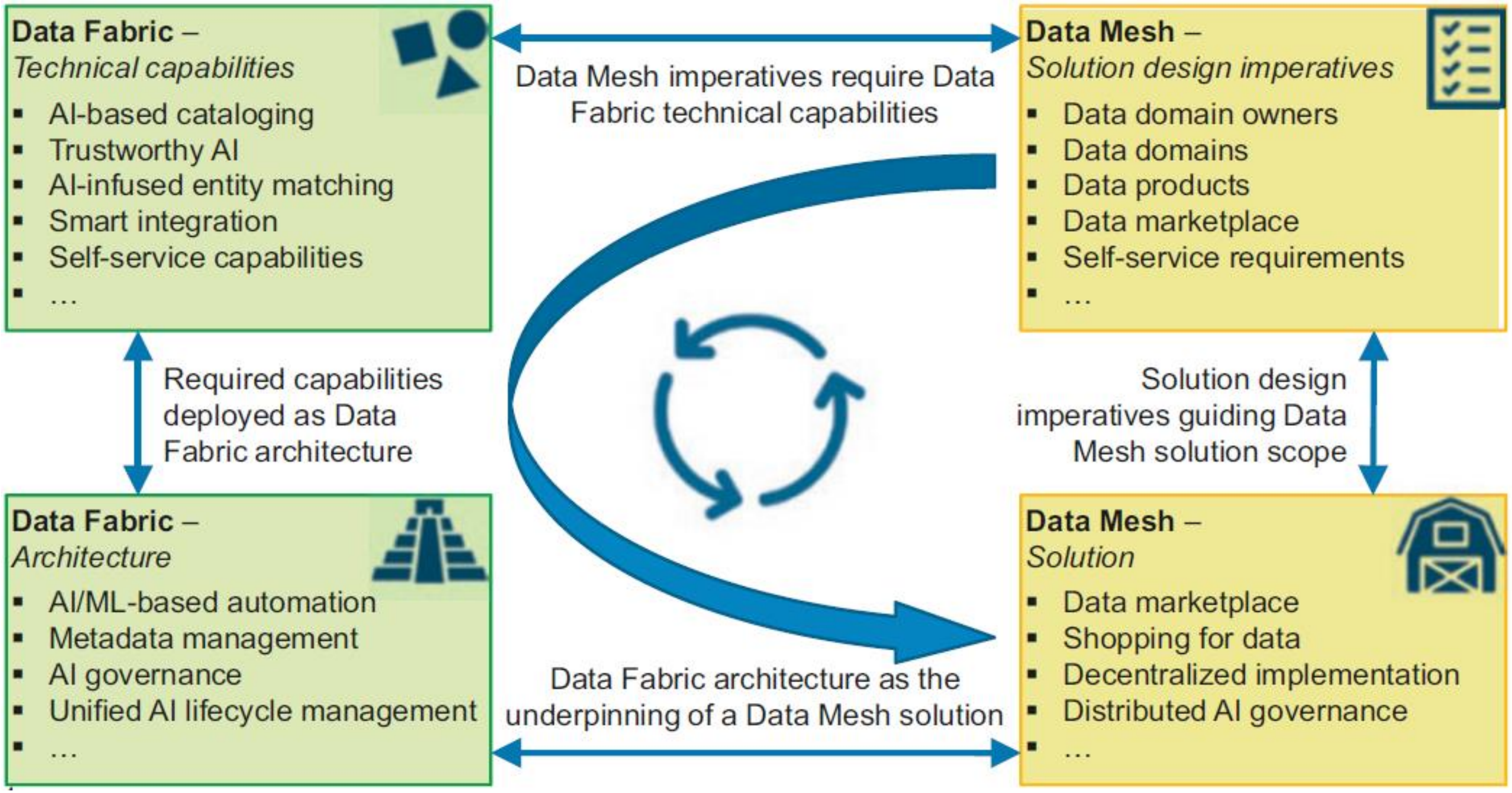
Data mesh vs data fabric

Both provide an architectural framework to access data across multiple technologies and platforms

- **Data fabric**
  - Attempts to centralize and coordinate data management
  - Tackles the complexity of data and metadata in a smart way that works well together
  - Focus on the architectural, technical capabilities, and intelligent analysis to produce active metadata supporting a smarter, AI-infused system to orchestrate various data integration styles
- **Data mesh**
  - Emphasis on decentralization and data domain autonomy
  - Focuses on organizational change; it is more about people and process
  - Data are primarily organized around domain owners who create business-focused data products, which can be aggregated and consumed across distributed consumers

Alex Woodie, 2021 https://www.datanami.com/2021/10/25/data-mesh-vs-data-fabric-understanding-the-differences/
Dave Wells, 2021 https://www.eckerson.com/articles/data-architecture-complex-vs-complicated

Data mesh vs data fabric



Data mesh vs data fabric

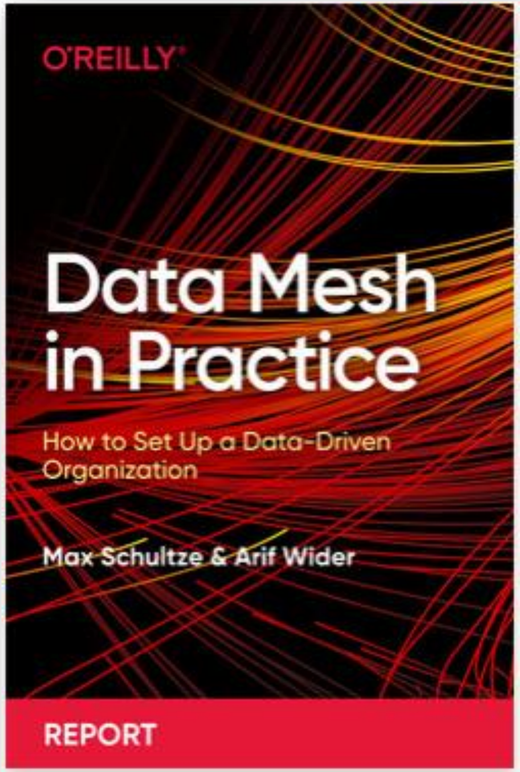
Data Fabric and Mesh are the results from the data architecture evolution

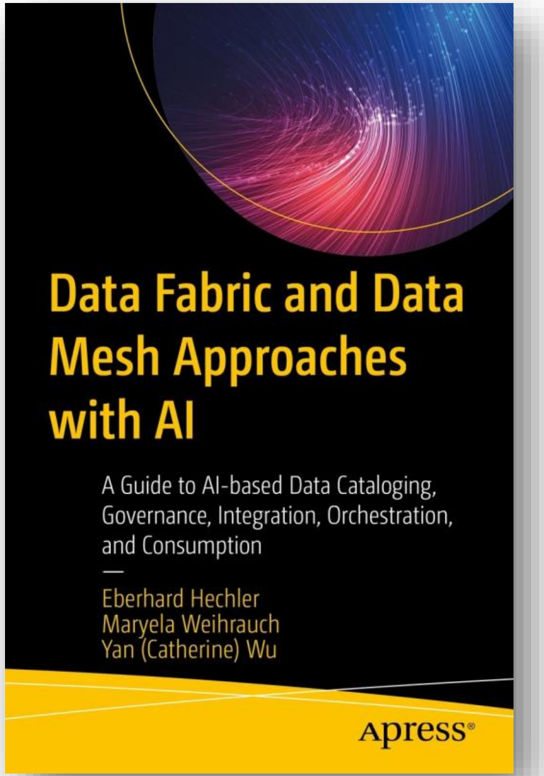
- **Many capabilities were in existence already long before** the terms were coined

Take away:

- Abstract the “building blocks” of such platforms
- Let them evolve according to scalability and flexibility requirements

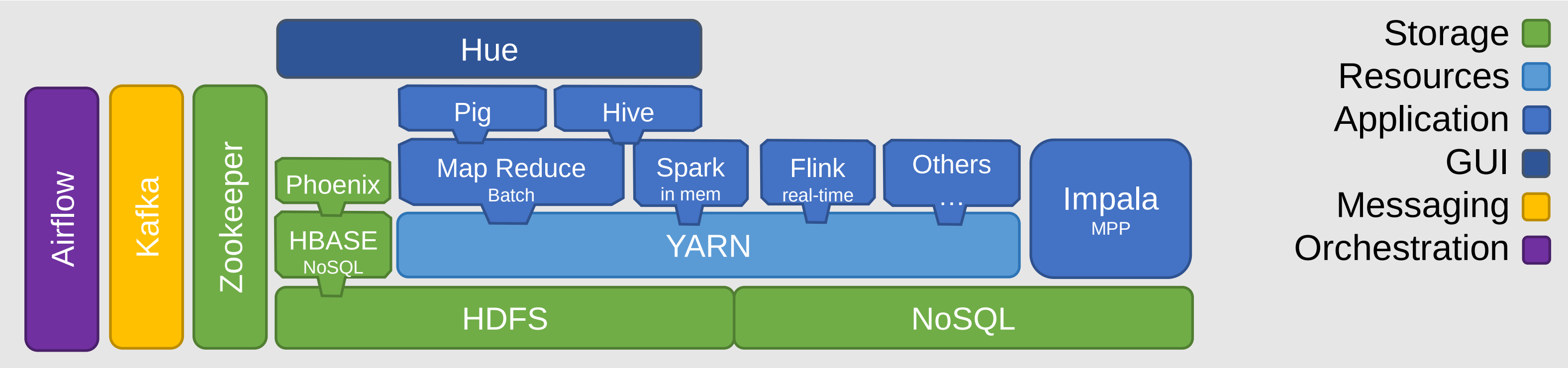
(Some) References





Example of data platform: Hadoop-based

A data platform on the Hadoop stack requires several tools



How many levels of complexity are hidden here?

How do you provision it?

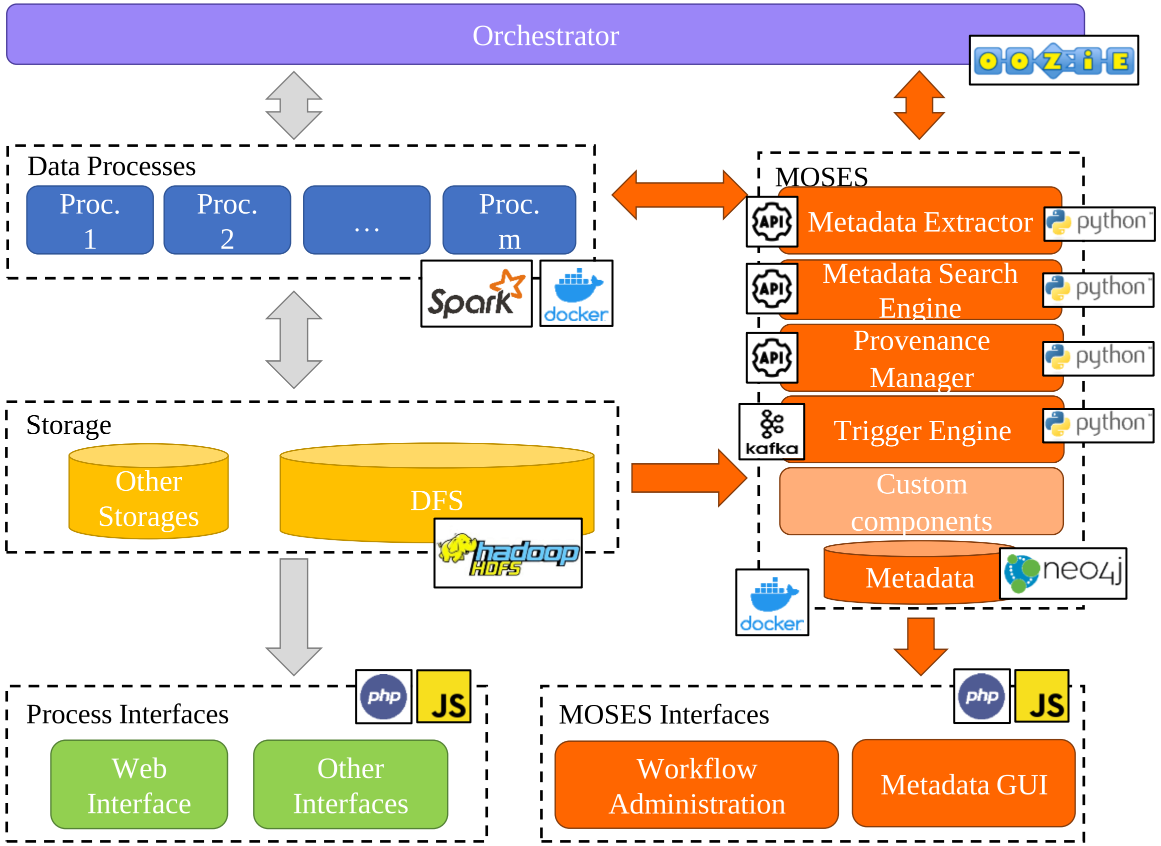
- Manual provisioning on-premises
- Semi-automatic provisioning on-premises
- Automatic provisioning in the cloud

Example of data platform: MOSES

Example of a data platform (MOSES)

Functional architecture

- Components of MOSES are in orange
- Others are standard components in charge of producing/consuming, processing, storing, and visualizing data
- The orchestrator (e.g., Oozie) manages (e.g., schedules) the data transformation processes



Francia, M., Gallinucci, E., Golfarelli, M., Rizzi, S. et al. (2021). Making data platforms smarter with MOSES. Future Generation Computer Systems, 125, 299-313.

Summing up

- Storage should be flexible enough to support heterogenous data models and raw data
  - From operational databases to DWHs **(why?)**
  - From relational data models to NoSQL **(why?)**
  - Data lake to (directly) ingest raw data
- Storage, *per se*, is insufficient to get value from the data **(examples?)**
  - We also need data processing and fruition
  - Data lakes are blurring into data platforms
- Data platforms support end-to-end data needs **(which ones?)**
  - Building data platforms is hard **(why?)**
  - Managing data platforms is hard, exploit meta-data to ease this task
    - Data lineage, compression, profiling, resolution, etc.
- **Open question**: how do we deploy working data platforms?