# Utilizing Data Lakes for Managing Multidisciplinary Research Data

Mark Greiner

Max-Planck Institute for Chemical Energy Conversion

2022-06-27

# Presentation Outline

Problem Space

- Description of domain
- Problem statement

Solution Space

- Data Governance
- Architectures
- Technologies
- Future directions

# 1. Description of the Domain

Problem space

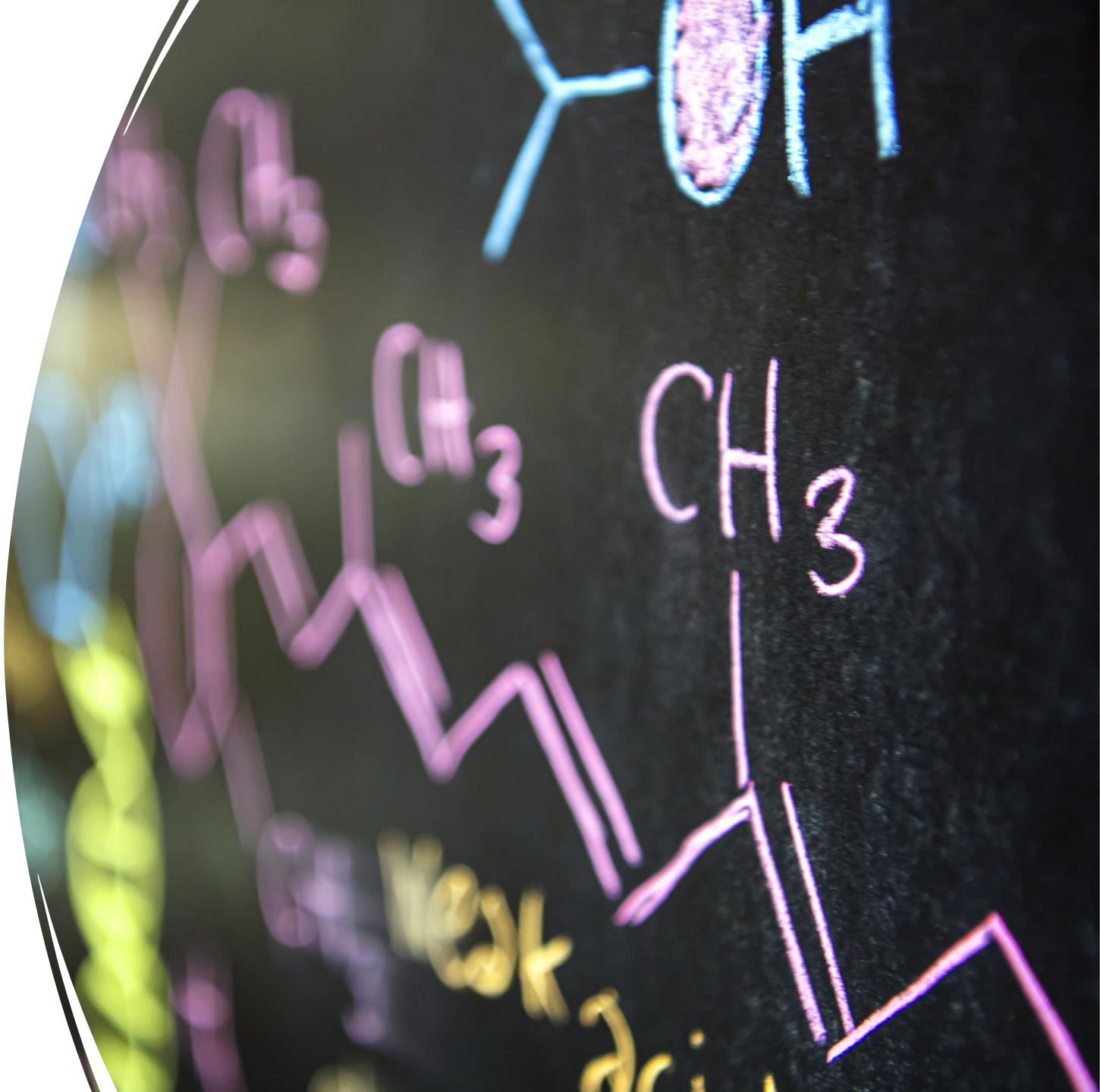
# About MPI-CEC

Research Discipline

- Catalysis for chemical energy
- Water splitting, bio-catalysts, chemical production

Magnitude

- 218 Researchers
- 22 Research Groups
- 3 Departments



# Diverse requirements

**Synthesis Lab**

- Focus on:
  - Creating new chemicals
- Workflow:
  - Plan, Synthesize, Characterize, Analyze results, Iterate, Test in some application
- Challenges:
  - Multi-disciplinary
  - Harmonize ELN with diverse data sources

# Diverse requirements

Self-service facilities

- Focus on:
  - routine measurements
- Can be performed with minimal training
- Workflow:
  - Schedule, Measure, Retrieve data, Analyze data,
- Challenge:
  - Harmonize users' ELN with instrument
  - Associate data with sample

# Diverse requirements

Testing facilities

- Focus on:
  - Behavior in applications
  - Testing parameters
- Workflow:
  - Plan, Schedule, Measure, Retrieve data, Analyze data
- Challenges:
  - Analysis
  - Linking data-sample-conditions

# Diverse requirements

Large facilities

- Focus on:
  - Characterizing
- Workflow:
  - Plan, Schedule, Measure, Retrieve data, Analyze data
- Major challenges:
  - Data sizes
  - Integrating with home ELN

# Roles and skills

## Student (Master/PhD)

### Skills

- Conducts experiments
- Documents results

### IT interactions

- Measure things
- Interact analysis software

## Post-Doc

### Skills

- Designing experiments
- Analysis workflows
- Documents results

### IT Interactions

- Interact measurement software
- Interact analysis software
- Supervise students
- Review results

## Principal investigator

### Skills

- Provides research questions
- Supervises research of Students and Post-Docs
- Administrative Tasks

### IT Interactions

- Interacts with management software

# Roles and skills



## Technical Staff

### Skills

- Maintenance Laboratory equipment
- Keep services running
- Administer stock
- Supervise experiments

### IT Interaction

- Interact with monitoring software



## Group Leaders

### Skills

- Project management

### IT Interaction

- Interact with management software

## Directors

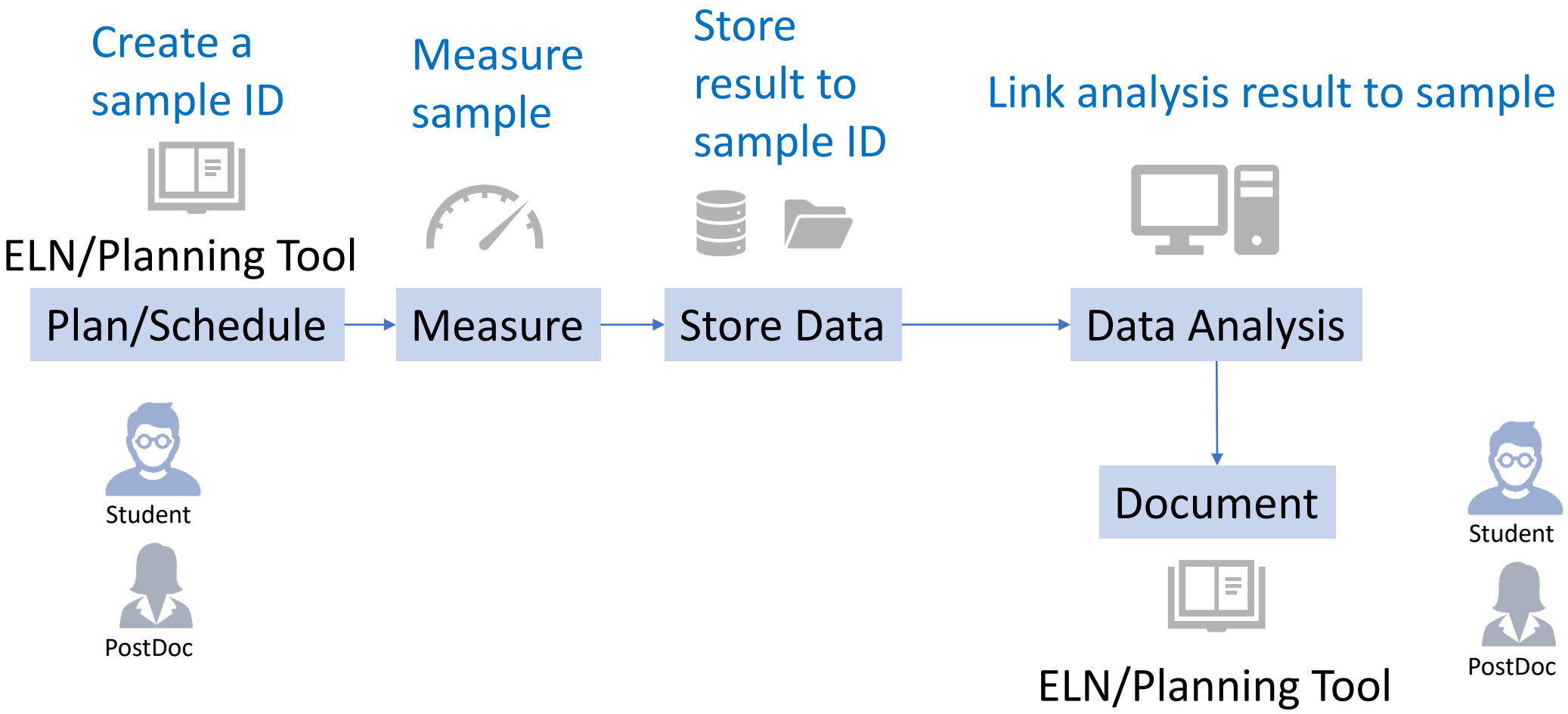
### Skills

- Management

### IT Interaction

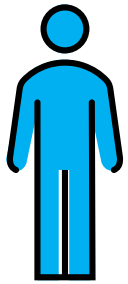
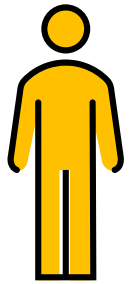
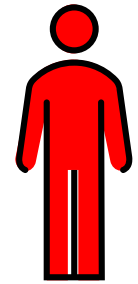
- View dashboards and reports

# Use Case: Self-service testing facitlity




Modified from 2021-05-31_NFDI4Cat Consortium meeting - ELN task force.pptx

# Different Users; Same Data

| Measurement Requester | Instrument Operator | Data Analyst |
| --- | --- | --- |
|  |  |  |
| Provides samples<br>Provides measurement requirements | Performs measurements<br>Provides data | Processes the data<br>Analyzes the data |
| ELN/LIMS | Instrument control software | Analysis software |
| Database | Semi-structured application format | Semi-structured application format |

# 2. Problem Statements

Problem Space

# Problem Statement

- Data assets are not organized.
- It is distributed across many locations, with no contextual metadata.
- Thus, searching and organizing tools cannot be used to utilize the data.
- Knowledge cannot be automatically extracted from it.

# Problem Statements

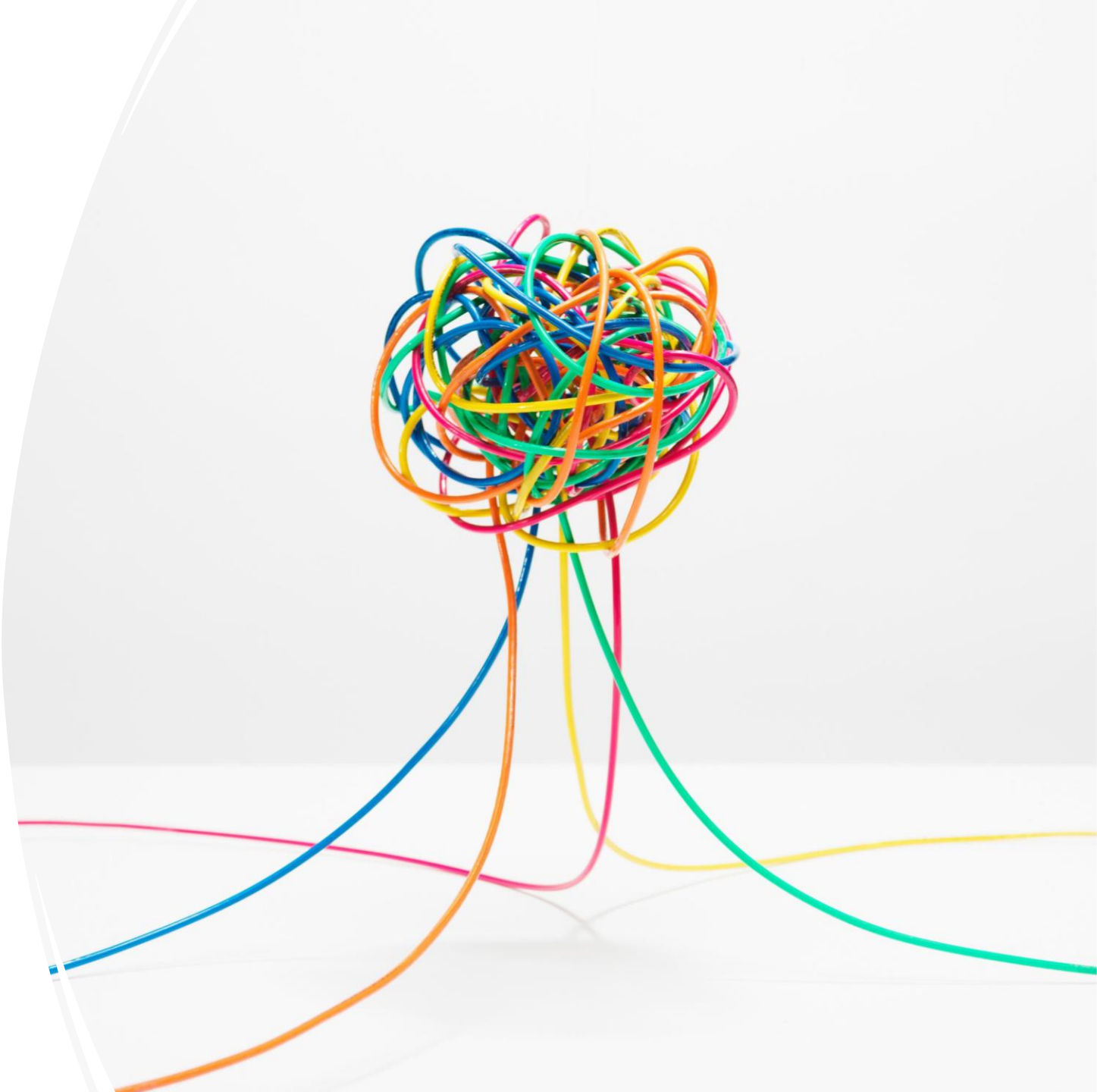
- Researchers spend too much time on repetitive manual work, related to organizing, searching and processing data.
- Takes away from value-added work, increases errors, leads to re-work.

# Problem Statements

- Users do not have easy access to all their data assets.

# Problem Statement

- It is difficult, sometimes impossible, to trace back the origin of a research result.
- Leads to excessive time spent searching when report revisions are needed.
- Decreases knowledge retention.



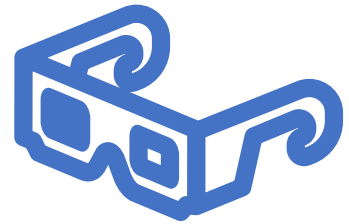
# Problem Statement

- The structure of the organization's data assets are not suitable for large-scale analysis algorithms.
- Unable to utilize modern algorithms for meta-analysis.

# Example of governance problems

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | building_name | building_identificati | building_identificatic | room_number | floor_numb | storage_name | storage_characteristic | |
| 2 | 1 | Physikgebaude | D | Physikgebaude | 110 | TKG | Kuhlschrank 2 +4Â°C | | |
| 3 | 4 | Physikgebaude | D | Physikgebaude | 102 | TKG | 102-1-Hazards | | |
| 4 | 5 | Physikgebaude | D | Physikgebaude | 102 | TKG | 102-2-Solvents | | |
| 5 | 6 | Physikgebaude | D | Physikgebaude | 103 | TKG | Chem | einfacher Schrank ohne Absaugung | |
| 6 | 7 | Physikgebaude | D | Physikgebaude | 104 | TKG | 104-1-Hazards | | |
| 7 | 8 | Physikgebaude | D | Physikgebaude | 104 | TKG | 104-3-Bases | | |
| 8 | 9 | Physikgebaude | D | Physikgebaude | 104 | TKG | 104-4-Acids | | |
| 9 | 10 | Physikgebaude | D | Physikgebaude | 104 | TKG | Betriebsmitteltonne | Spannringfass 30L oder 60L zur Entsorgung | |
| 10 | 11 | Physikgebaude | D | Physikgebaude | 107 | TKG | 107-3-Hazards | | |
| 11 | 12 | Physikgebaude | D | Physikgebaude | 107 | TKG | 107-4-Hazards | | |
| 12 | 13 | Physikgebaude | D | Physikgebaude | 107 | TKG | 107-1-Solvents | | |
| 13 | 14 | Physikgebaude | D | Physikgebaude | 107 | TKG | Kuhlschrank | | |
| 14 | 15 | Physikgebaude | D | Physikgebaude | 107 | TKG | 107-5-Acids | | |
| 15 | 16 | Physikgebaude | D | Physikgebaude | 107 | TKG | Betriebsmitteltonne | Spannringfass 30L oder 60L zur Entsorgung | |
| 16 | 17 | Physikgebaude | D | Physikgebaude | 110 | TKG | 110-2-Hazards | | |
| 17 | 18 | Physikgebaude | D | Physikgebaude | 110 | TKG | 110-1-Inorg Acids | | |
| 18 | 19 | Physikgebaude | D | Physikgebaude | 110 | TKG | Betriebsmitteltonne | Spannringfass 30L oder 60L zur Entsorgung | |
| 19 | 20 | Physikgebaude | D | Physikgebaude | 111 | TKG | Glovebox A | nur fur ungefÃ¤hrliche Substanzen geeignet | |
| 20 | 21 | Physikgebaude | D | Physikgebaude | 111 | TKG | Glovebox B | nur fur ungefÃ¤hrliche Substanzen geeignet | |
| 21 | 22 | Physikgebaude | D | Physikgebaude | 101 | TKG | Glovebox C | nur fur ungefÃ¤hrliche Substanzen geeignet | |
| 22 | 24 | Physikgebaude | D | Physikgebaude | 113 | TKG | F90 Gase | F90 | |
| 23 | 25 | Physikgebaude | D | Physikgebaude | 113 | TKG | F90 Gase Formiergas | F90 | |
| 24 | 26 | Physikgebaude | D | Physikgebaude | 113 | TKG | Betriebsmitteltonne | Spannringfass 30L oder 60L zur Entsorgung | |
| 25 | 27 | Physikgebaude | D | Physikgebaude | 142 | TKG | Hausdienste | ganzer Raum | |

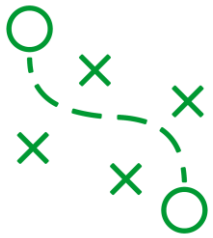
# 3. Data Governance



Solution Space

# What is a data strategy?

- A data strategy can be considered as an approach that enables to derive new knowledge from data
- A comprehensive data strategy is the basis for the successful implementation of data related projects



A Data Strategy describes the …

- … organizational structure for the successful use of data with relevant processes for dealing with data
- … required skills & roles
- … technology and tools

https://www.gartner.com/en/documents/3975595/creating-a-data-strategy

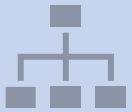
# Goals of a Data Strategy

Remove Information Silos

Keep knowledge of research

Make reasearch more accessible

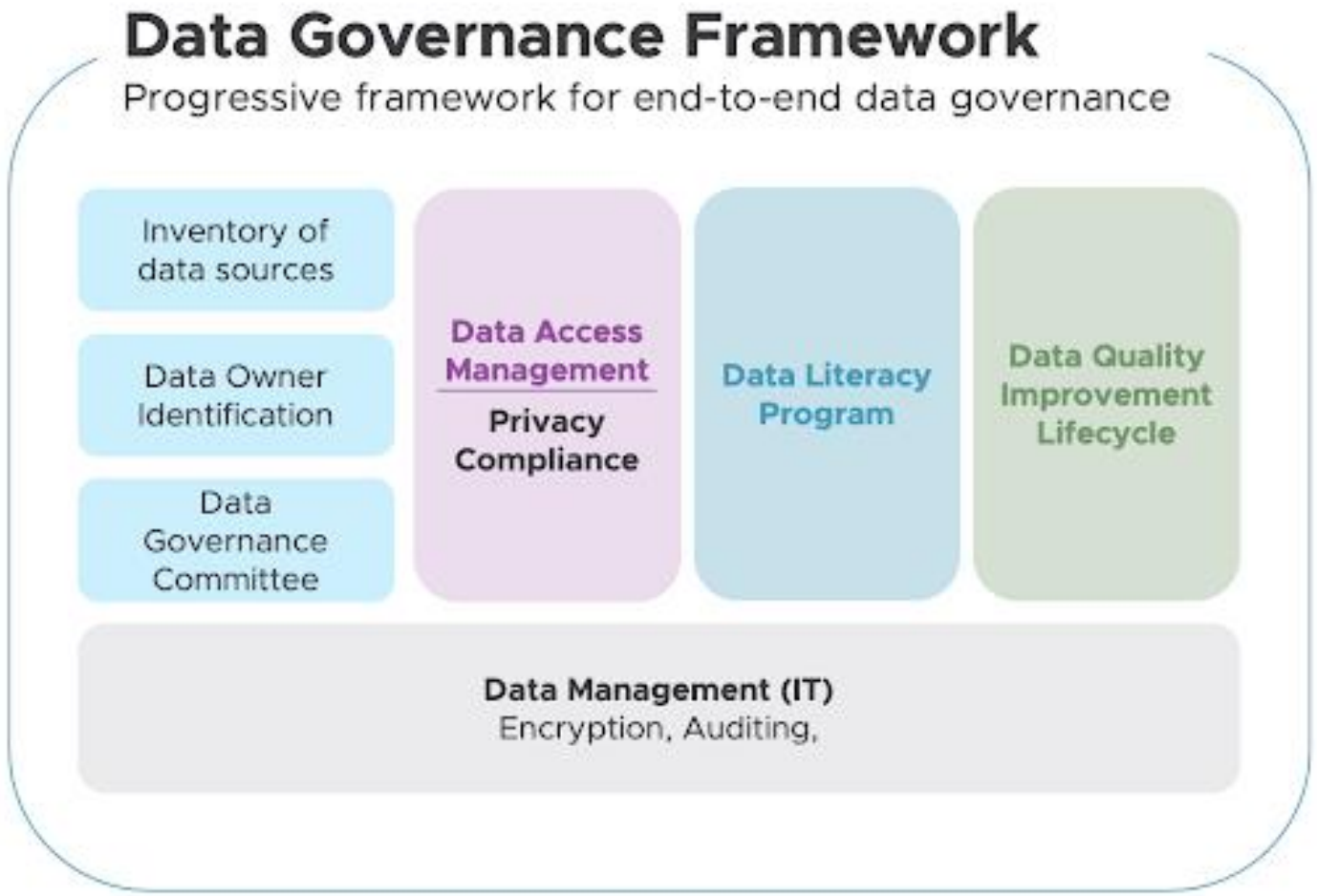
Enable meta research



Improve sustainability

https://datagovernance.com/goals-and-principles-for-data-governance/

# Data Governance Concepts




[Building an Effective Data Governance Framework (ovaledge.com)]

# Roles and skills in a business context

## Data Engineer

Transform and harmonise data

## Data Architect

Provide data processing concepts

## Data Scientist

Analyse and model data

## Data Artist

Visualize data

## Data Custodian

Data storage and security

## Data Steward

Steering and household data

## Data Security Admin.

Data security concepts

## Domain Expert

Domain knowledge

## Data Evangelist

Explores data potential

# Data Governance Roles

## Student (Master/PhD)

DG Roles

- Data engineer
- Data scientist

## Post-Doc

DG Roles

- Data scientist
- Data architect
- Data engineer
- Domain expert

## Principal investigator

DG Roles

- Data steward
- Data evangelist
- Domain expert

# Data Governance Roles



## Technical Staff

DG Roles

- Data steward



## Group Leaders

DG Roles

- Data steward
- Data evangelist
- Domain expert

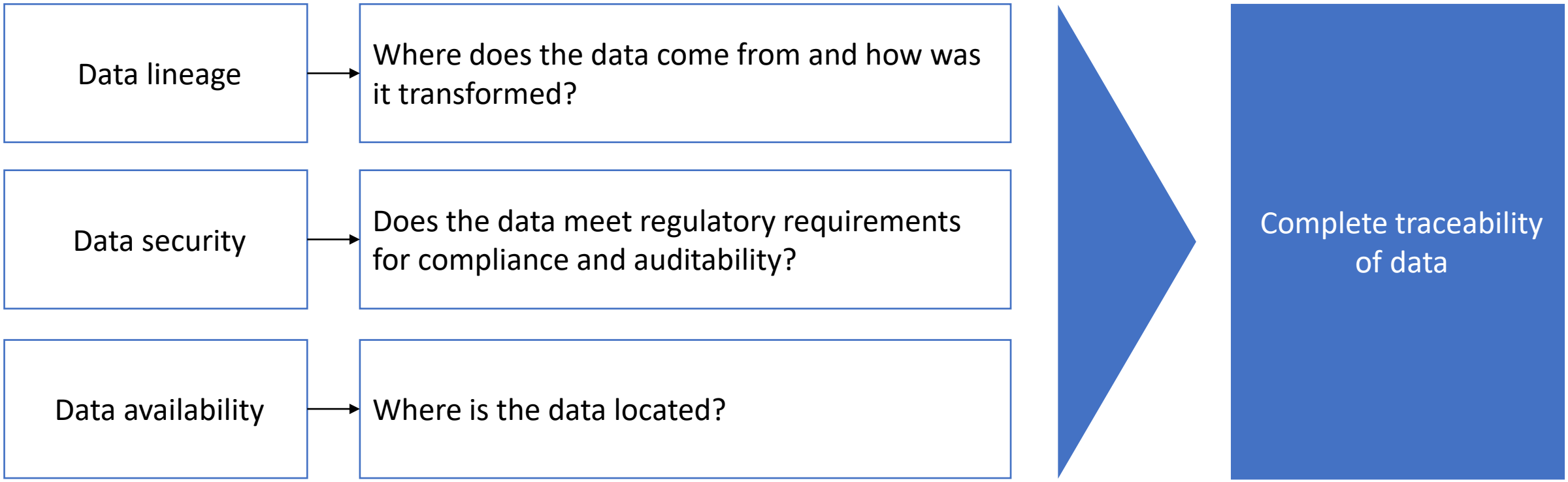
# Take-away message

- Data governance has specialized roles
- Academia often fulfills these roles by loading responsibility onto operational staff

- It may be time to have dedicated data-governance roles in academia

- However, due to scale and diversity at academia organizations, it may need a different structure than is used in industry

# Data Governance Concepts




https://datagovernance.com/goals-and-principles-for-data-governance/

# What is metadata?

- Data about data 
- Metadata can describe a single piece of data, a dataset or collection.
- Standard types of metadata:
  - *Descriptive*: information about **who** created a resource, **what** it is **about** and **what** it **includes** (e.g. title, author, subjects, keywords etc.)
  - *Structural*: information about the **way** data elements are **organized**, their **relationship** and the **structure** they exist in (e.g. ER-model)
  - Administrative: information about the origin of resources, their type and access rights
    (e.g. file type, date of creation etc.)

Table with **4** books, created by Joe Dow

Descriptive METADATA

| ID | Title | Author | Year | Cover | Edition | Price |
|---|---|---|---|---|---|---|
| 1 | Emma | Austen | 1815 | Paperback | 20th | € 4,30 |
| 2 | Dracula | Stoker | 1897 | Hardback | 15th | € 10,00 |
| 3 | Ivanhoe | Scott | 1820 | Hardback | 8th | € 20,00 |
| 4 | Kidnapped | Stevenson | 1886 | Paperback | 11th | € 3,50 |

**Origin of resources:** Book store
**Access rights**: read only - everyone; write and read – Joe Dow
**Created** on 5. January 2019

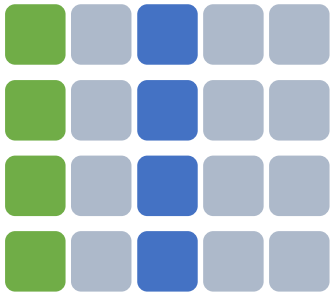
Administrative METADATA

Unsorted table;
related to sales data;
key is 'ID'

Structural METADATA

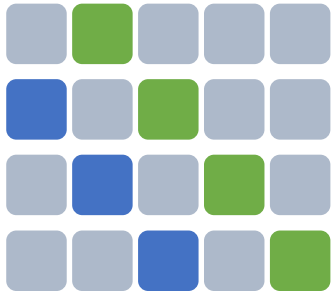
# What is data structure?

Data structure is the particular way of ORGANIZING & STORING digital information.



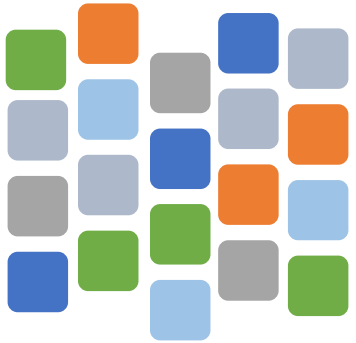
## Structured data:

Information with a specific and high degree of organization (tabular form)

## Semi-structured data:

Information with some degree of organization

## Unstructured data:

Information with no pre-defined organizational structure

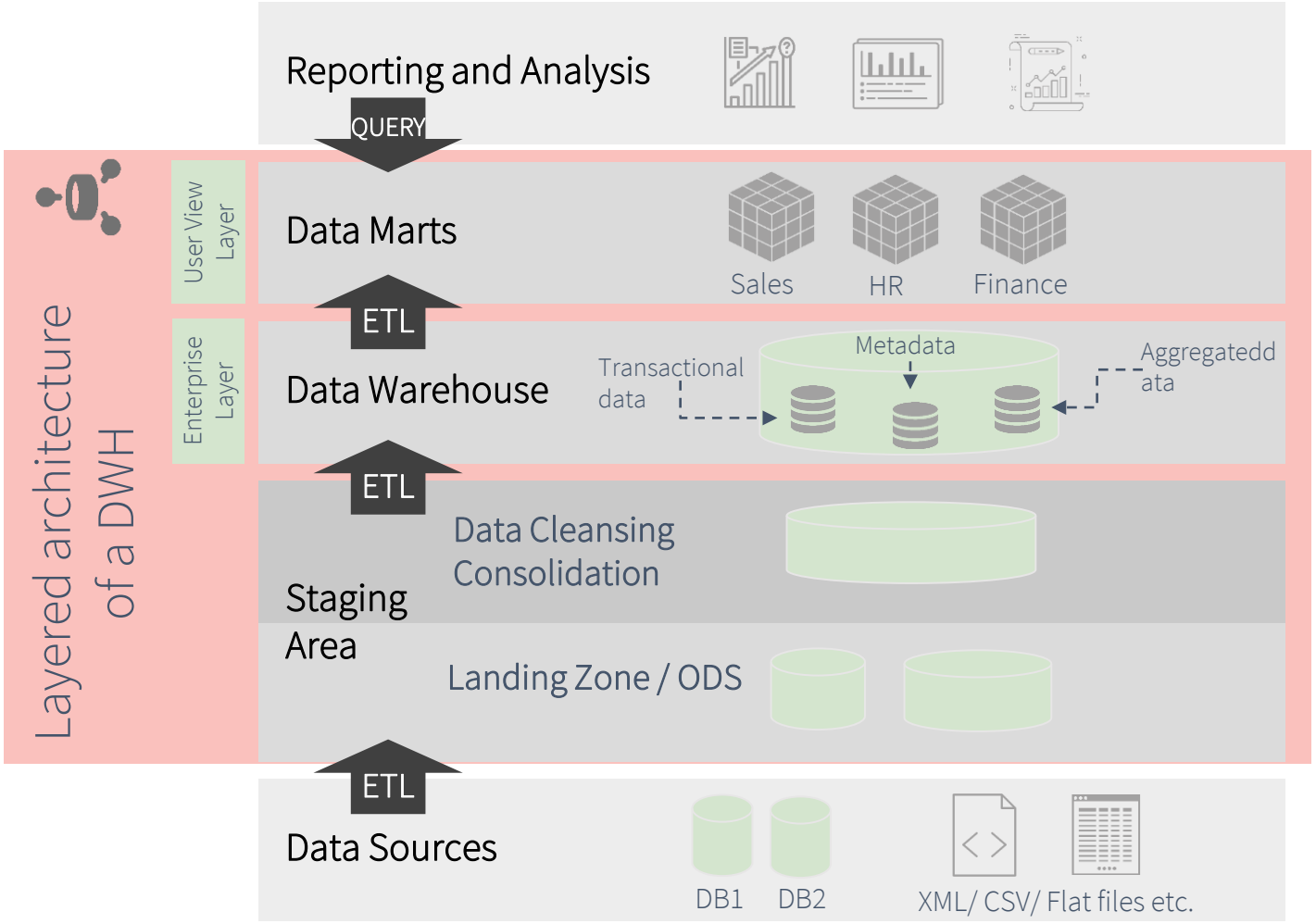
# How do we store data?

| | Databases | | File System | Object Storage |
|---|---|---|---|---|
| Definition | **Collection of stored data** organized in a specific way and determined by the data model underlying the database (e.g. ER-diagram)  | | • Store data in a space with a pre-defined scheme (e.g. a file hierarchy)  | • Store data as objects in a space with no pre-defined scheme<br>• Every file has a unique identifier, so it can be found (e.g. URL)  |
| | **Relational DB** | **Non-relational DB** | | |
| | • Based on relational data model (data is organized in tables)<br>• Structured Query Language (SQL) for querying the database | • Based on any data model other than the relational model<br>• Examples include key-value stores, document stores and graph databases | | |
| E.g. | • MS SQL Server, MySQL, Oracle | • MongoDB, Cassandra, Neo4J | • NTFS (Windows), Hadoop Distributed File System, ext3 | • AWS S3 buckets<br>• Azure Blob Storage |

# 4. Data Architectures

Solution Space

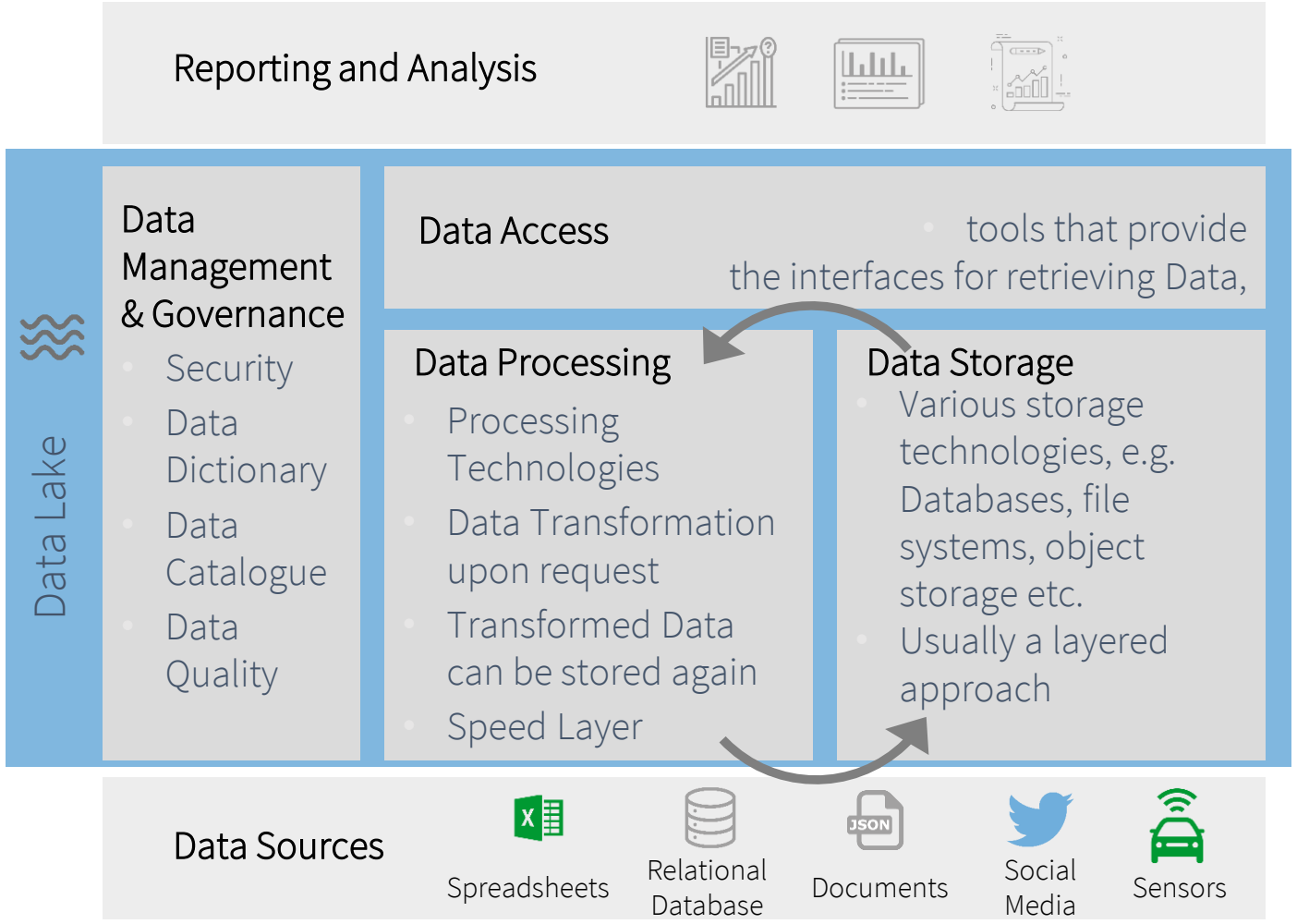
# What is a Data Warehouse?




"A Data Warehouse is a subject oriented, integrated, non-volatile and time-variant collection of data in support of management decisions"*

*Source: Immon, WH. Building the Data Warehouse

# What is a Data Lake?

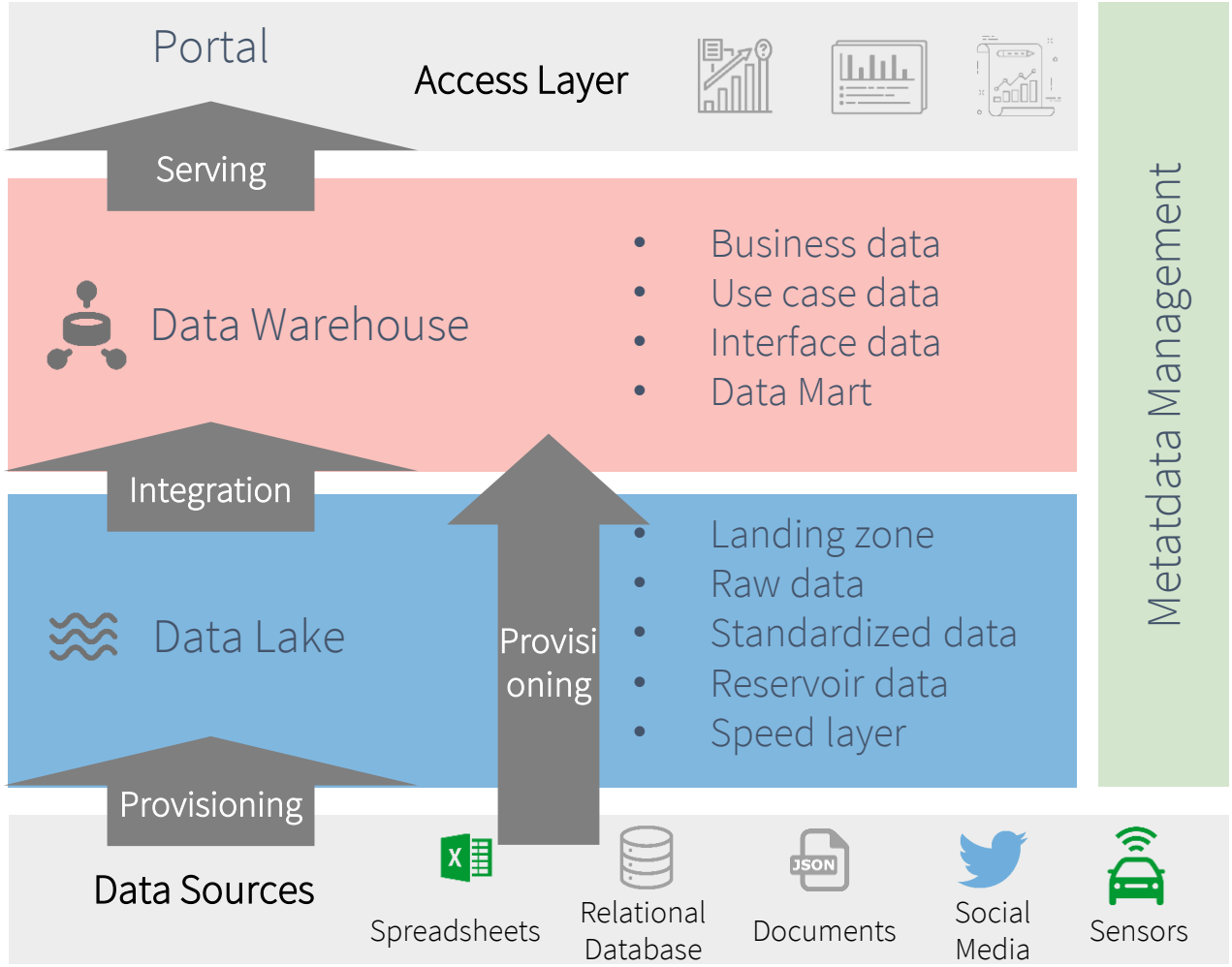



A Data Lake is a modular system of data storage and processing technologies. Like a DWH it is a logical concept rather than a tangible entity.

### Benefits of a Data Lake

- stores all kinds of data, e.g.:
  - Structured Tables
  - Text Documents
  - Pictures
- Central entry point for data access

# How can both architectures be combined?

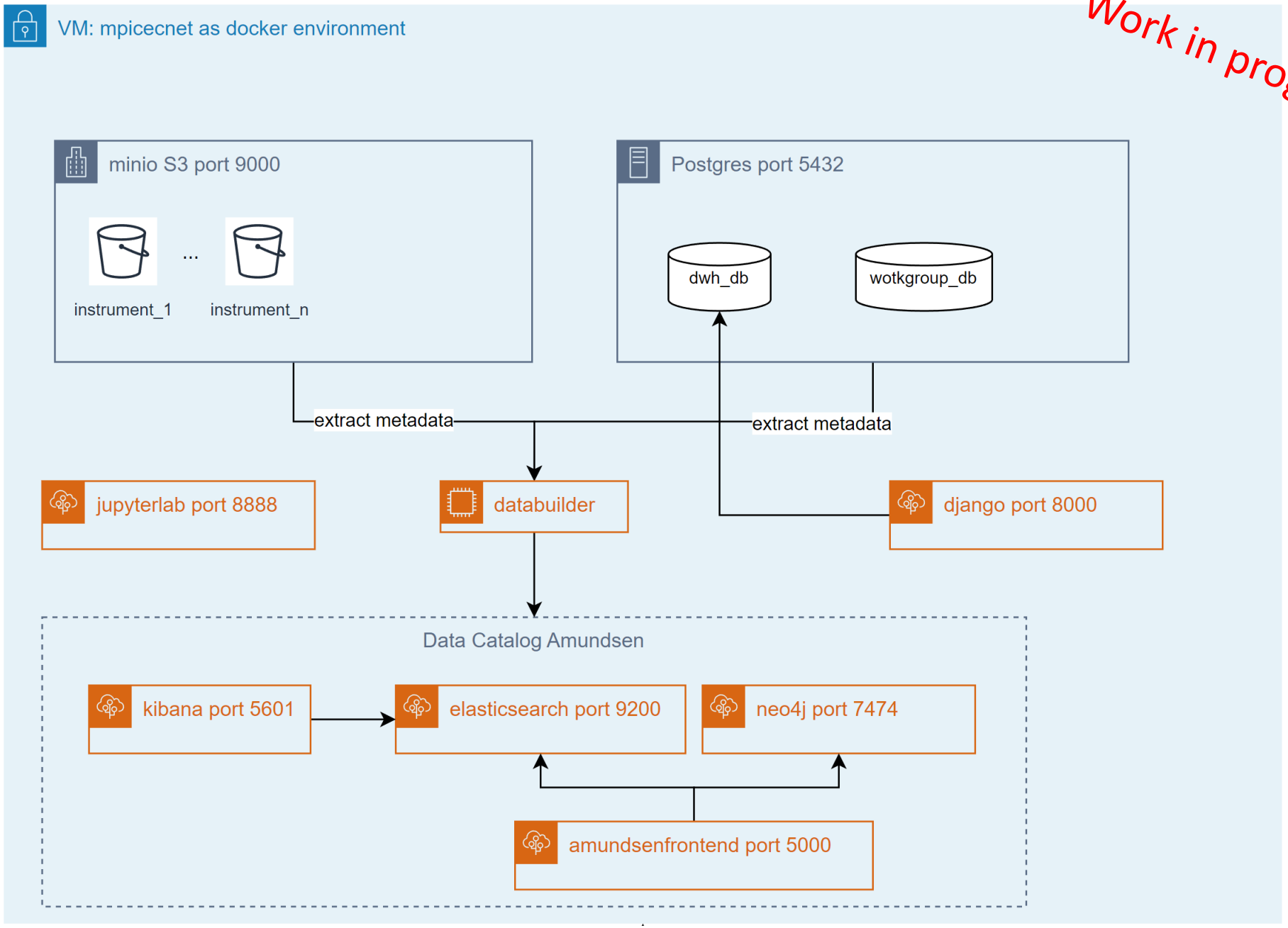



## Advantage

- One platform for different requirements
- Best of both architectures

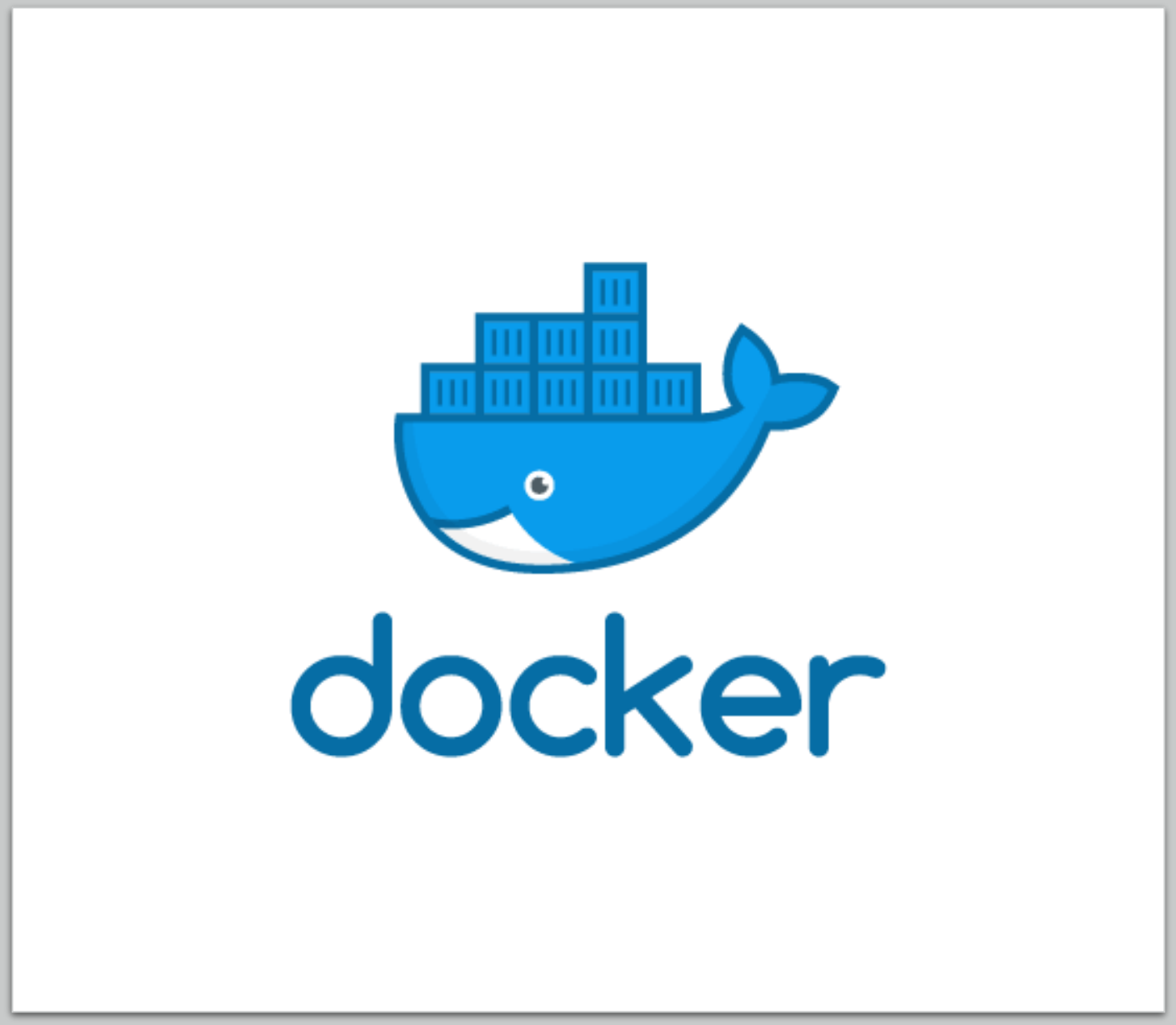
# 5. Technologies

Solution Space


VM: mpicecnet as docker environment
Work in progress
minio S3 port 9000
instrument_1
...
instrument_n
Postgres port 5432
dwh_db
wotkgroup_db
extract metadata
extract metadata
jupyterlab port 8888
databuilder
django port 8000
Data Catalog Amundsen
kibana port 5601
elasticsearch port 9200
neo4j port 7474
amundsenfrontend port 5000

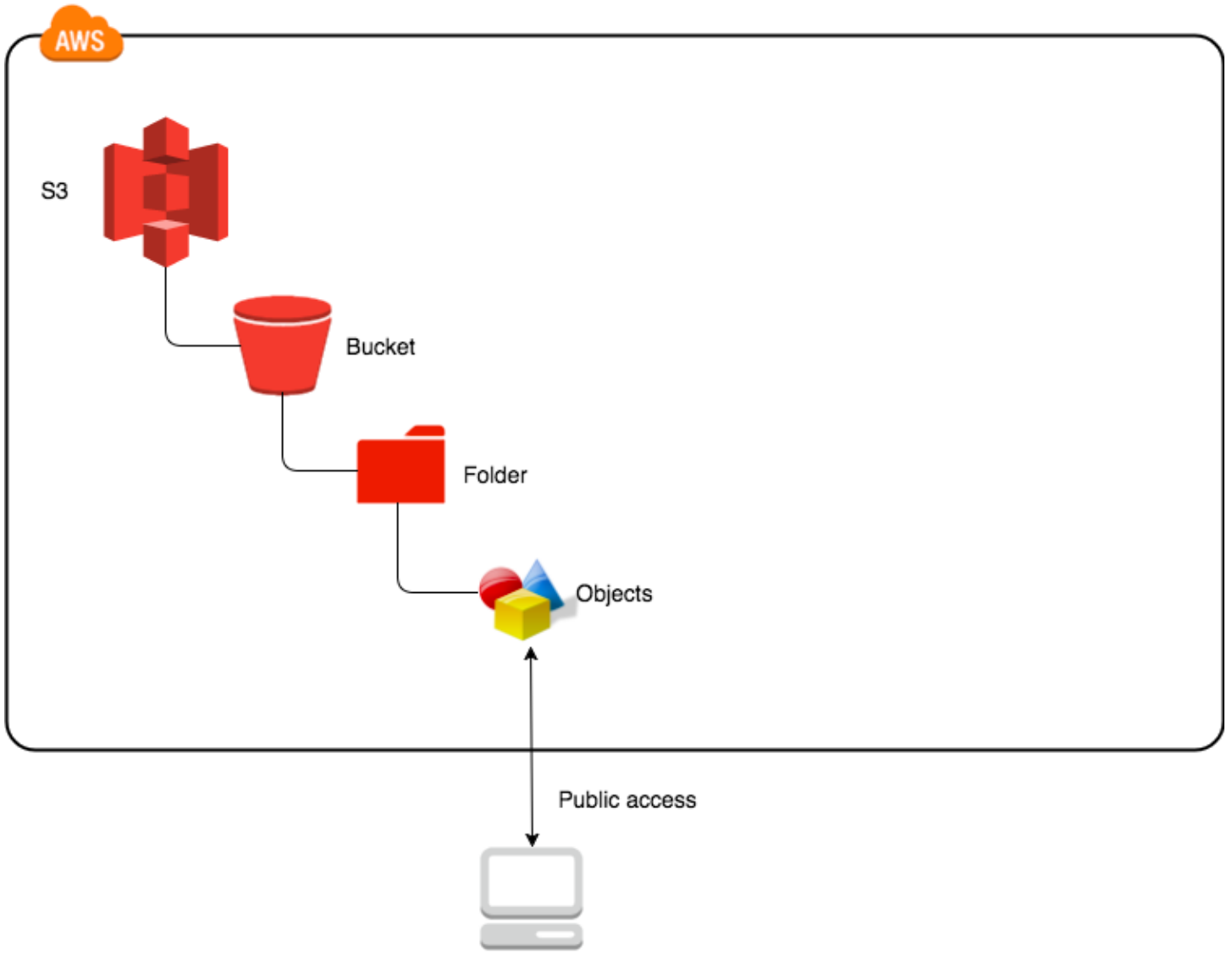
# Docker

- Containers for running microservices
- Isolated environments
- Needs less resources than a VM



# Minio

- Object storage
- Same as Amazon S3
- Can store any kind of data



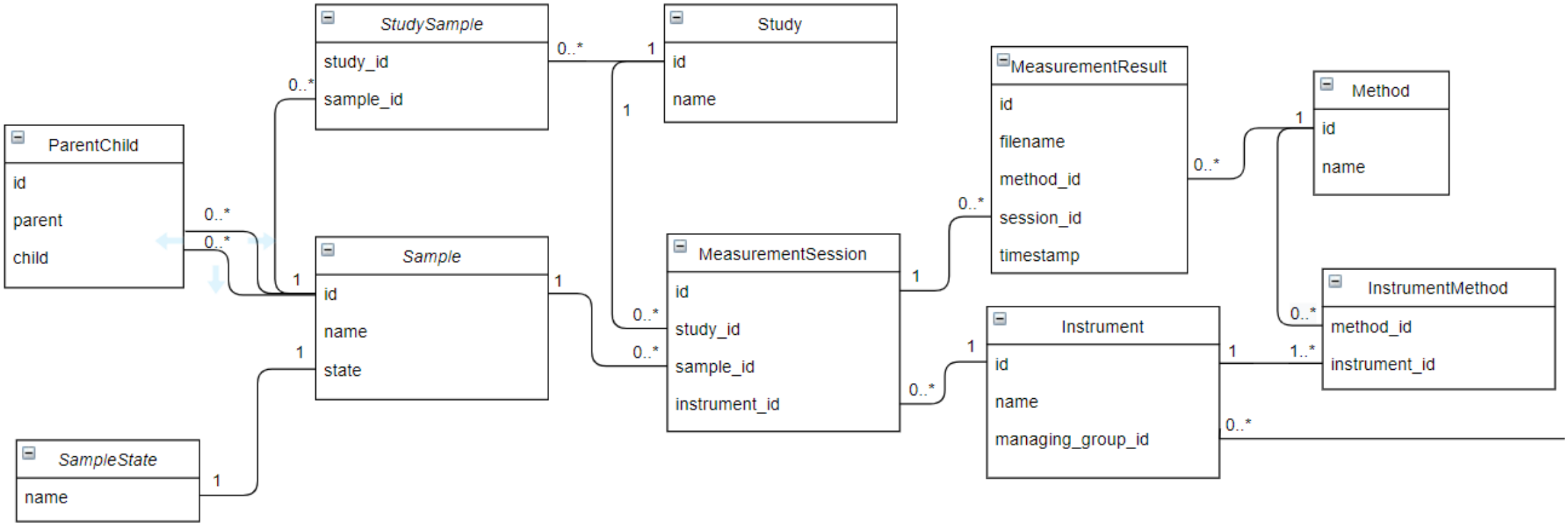

MINIO

- Buckets hold objects
  - Define access rights
- Objects have immutable metadata
- Client can access data over API

# PostgreSQL

- Relational database
- Used for storing structured data
- Transactional data and data warehouse
- Strict and robust data models

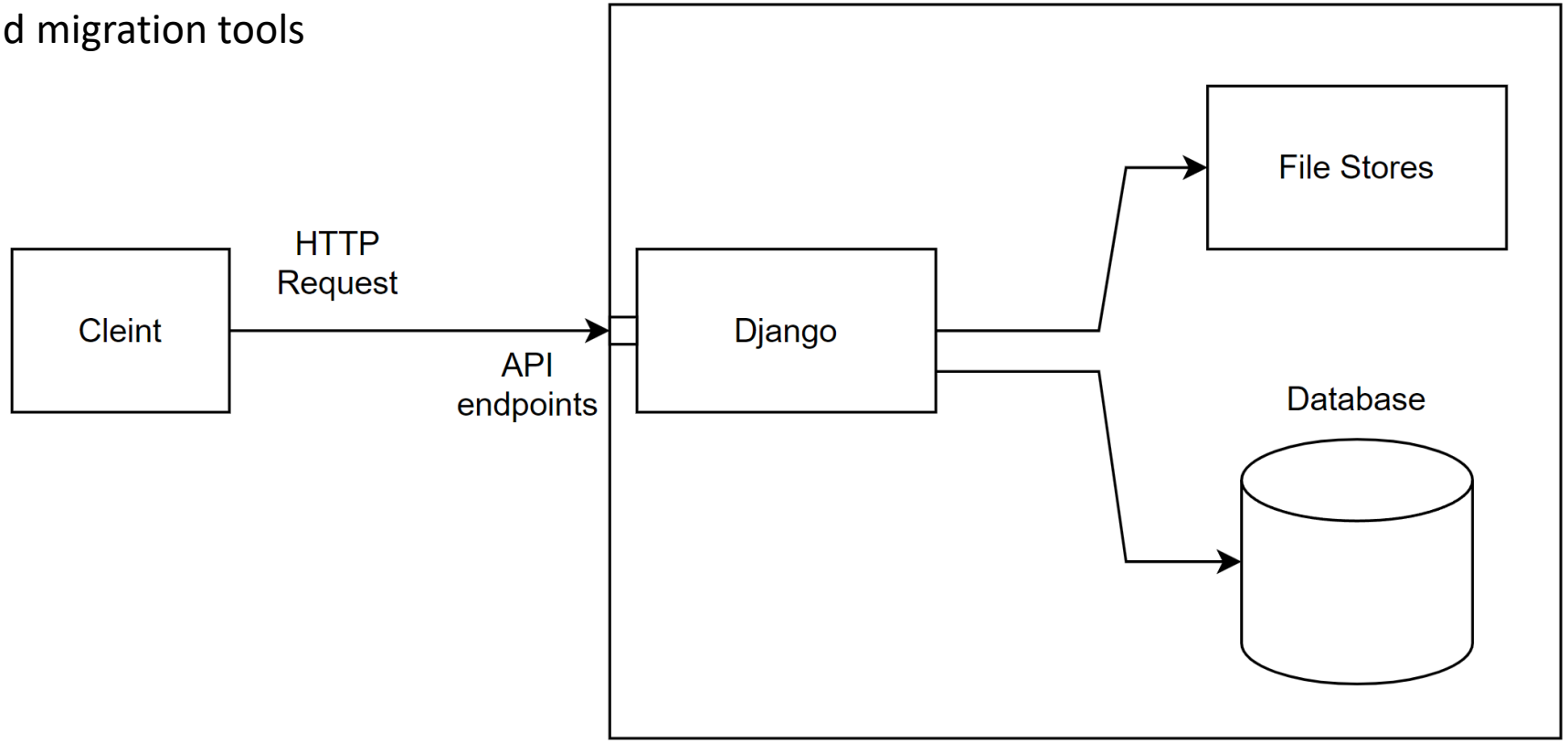
- Used for storing:
  - Transactional Data
  - Data Warehouse tables



# Django

- Back-end Web application framework written in python
- Robust and scalable
- Used for providing REST API for accessing Relational Database

- API endpoints provide client with ability to CRUD database entries
- Will be used to connect front-end apps
- Access to File stores
- Django provides Authentication and Authorization to access resources
- Django provides convenient database schema history and migration tools



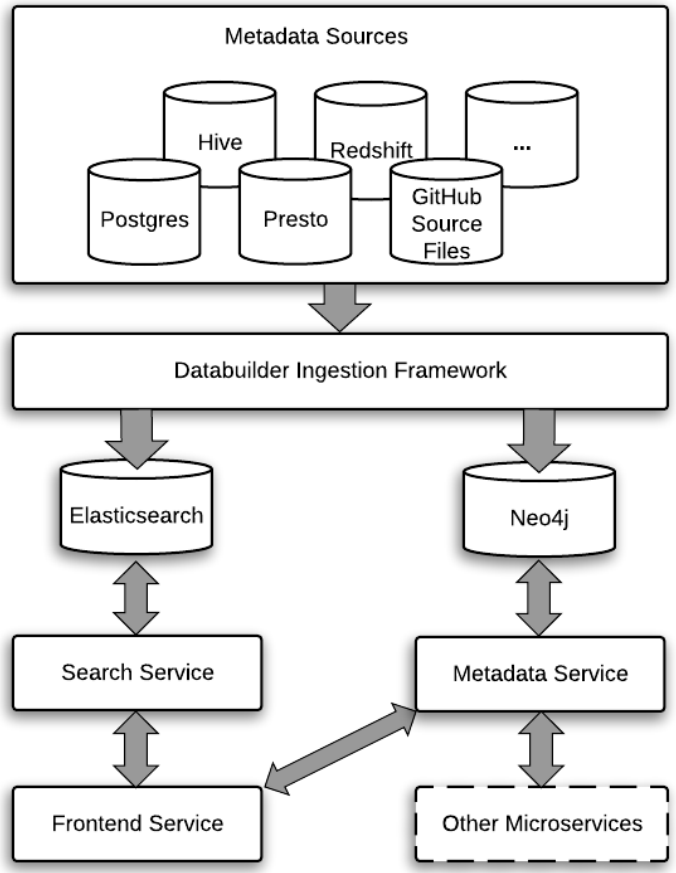
# Amundsen

- Data catalog
- PageRank-inspired search algorithm
- Provides REST API for search engine

# How does the data catalog work?

- Amundsen collects Metadata via the data builder ingestion framework
- Metadata and lineage data are stored in Neo4j and elasticsearch
- Metadata is made accessible via search interface
- Metadata can be made searchable for all users, whereas the content remains visible only with sufficient access permissions

## Amundsen data catalog architecture



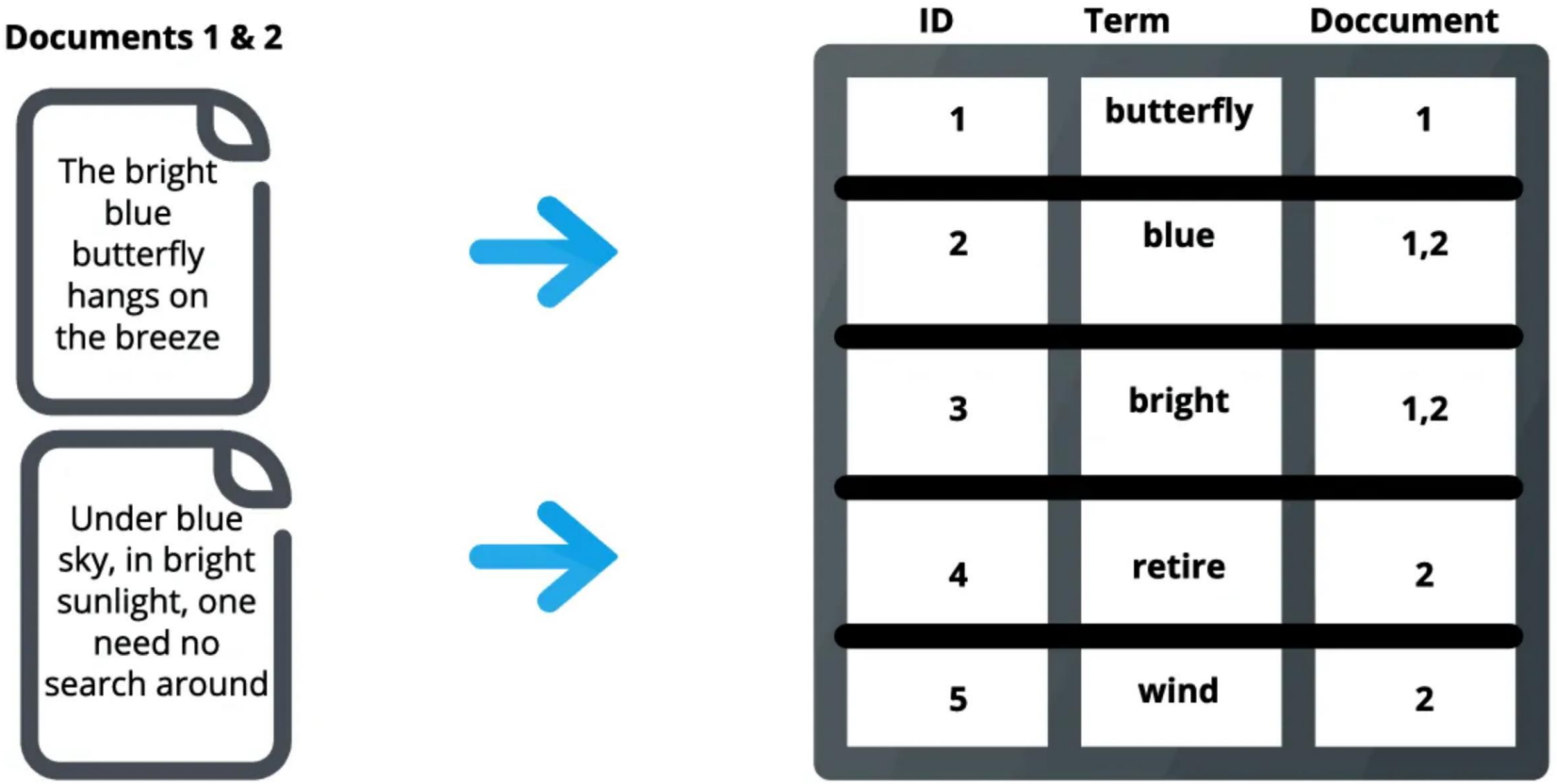

https://www.amundsen.io/amundsen/architecture/

# Elasticsearch

- Search engine
- Service behind Amundsen's search library
- NoSQL data store
- Stores data as documents (like JSON)

# Inverted Index

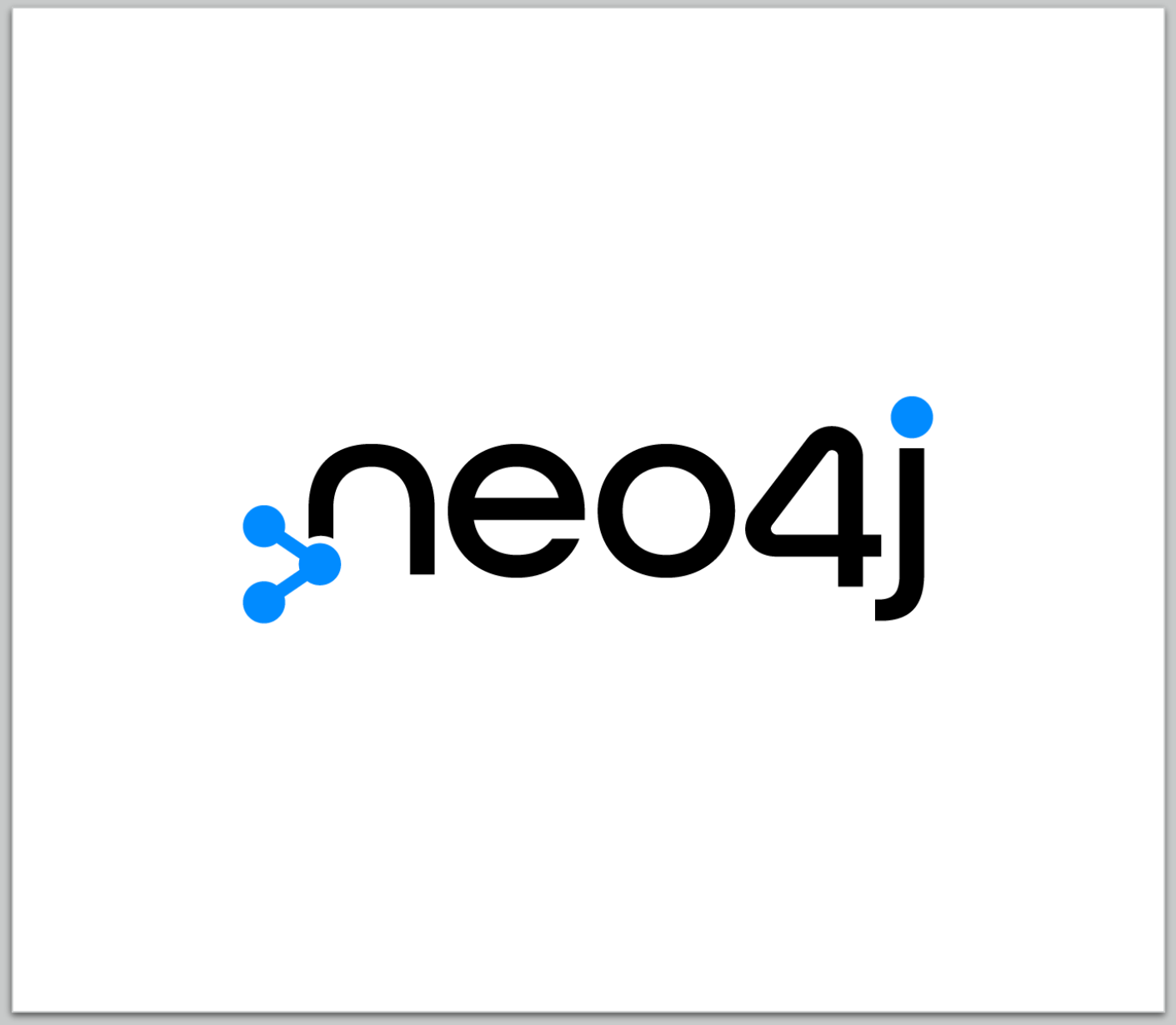
elasticsearch



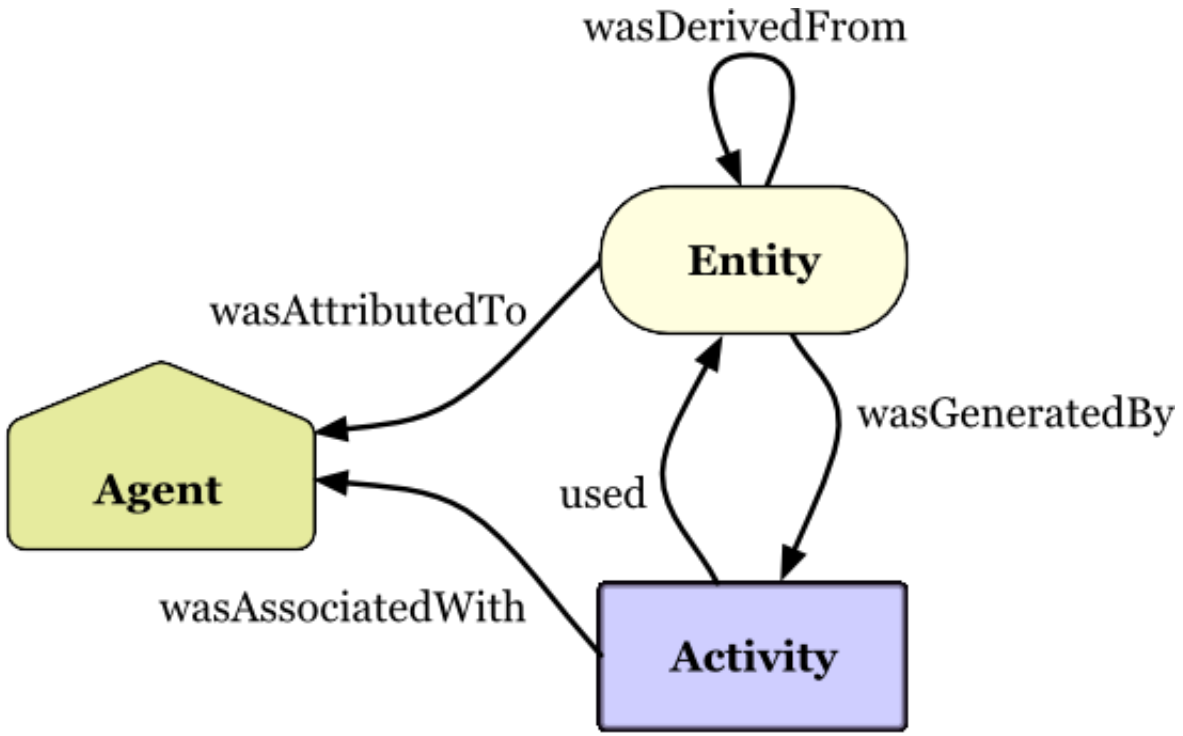

https://www.knowi.com/blog/what-is-elastic-search/

# Neo4j

- Graph database behind Amunden's metadata service
- Stores 'Triples' (subject, predicate, object)
- Great for traversing relationships



# Data Provenance

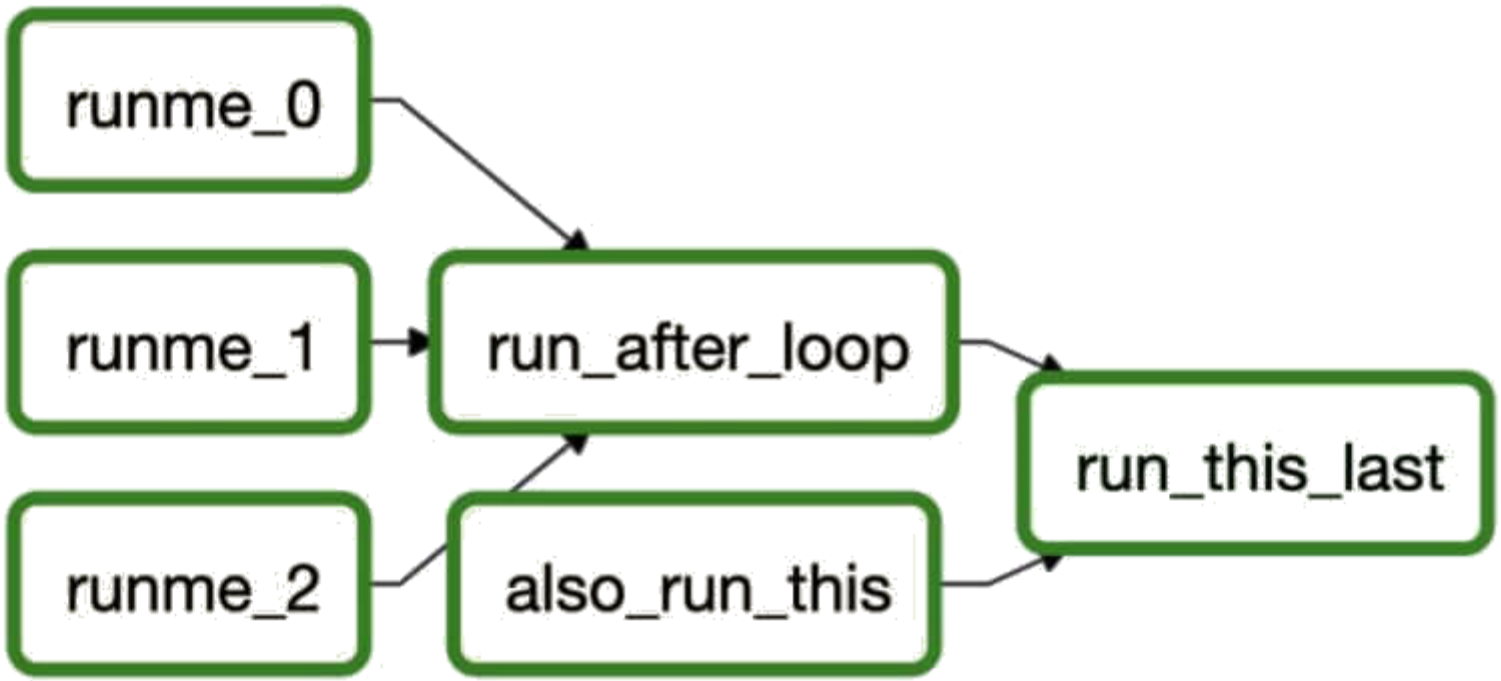



https://www.w3.org/TR/prov-primer/

# Apache Airflow

- Orchestration tool
- Build for running complex cron jobs
- Will be used for ETL processes

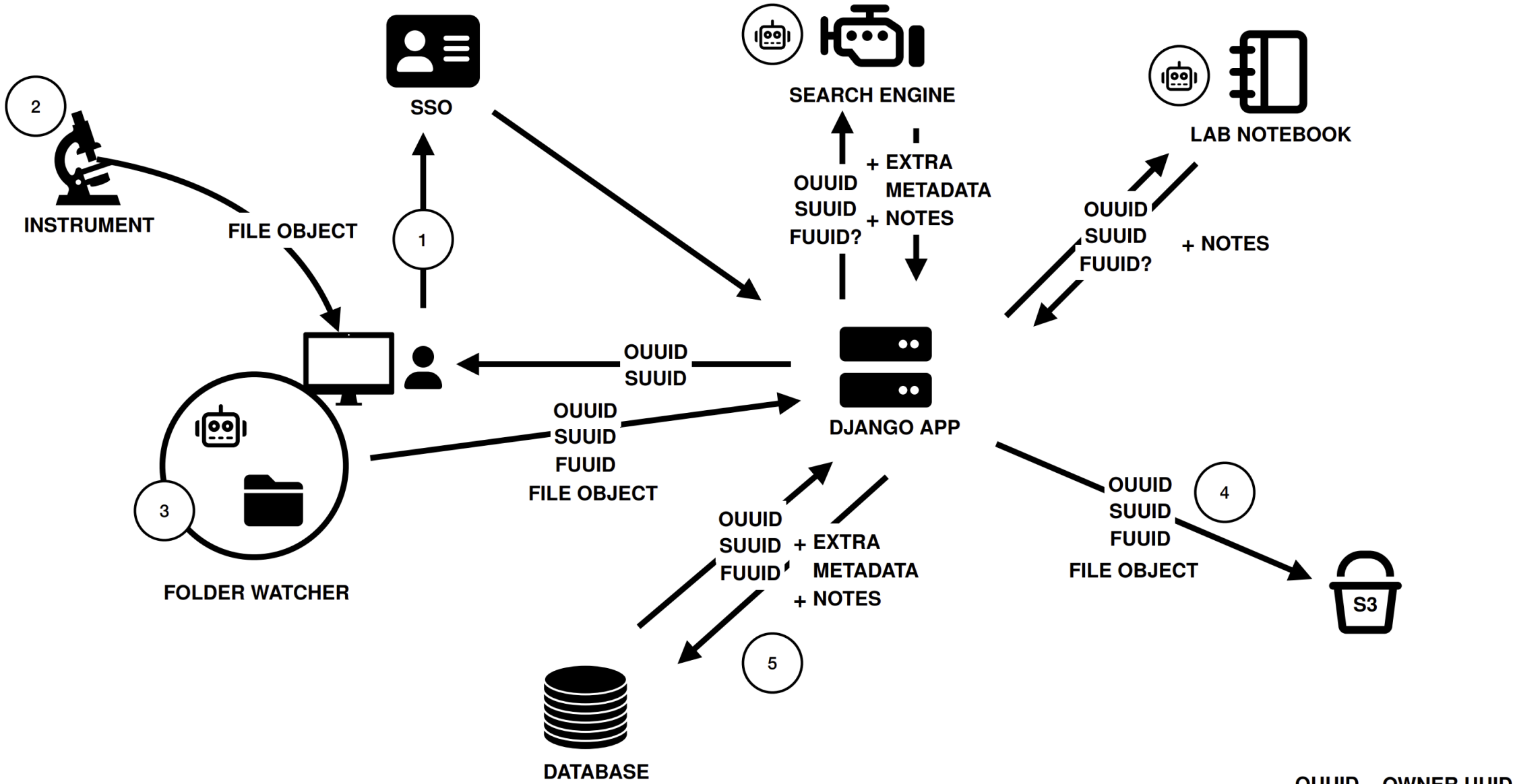
# Orchestration



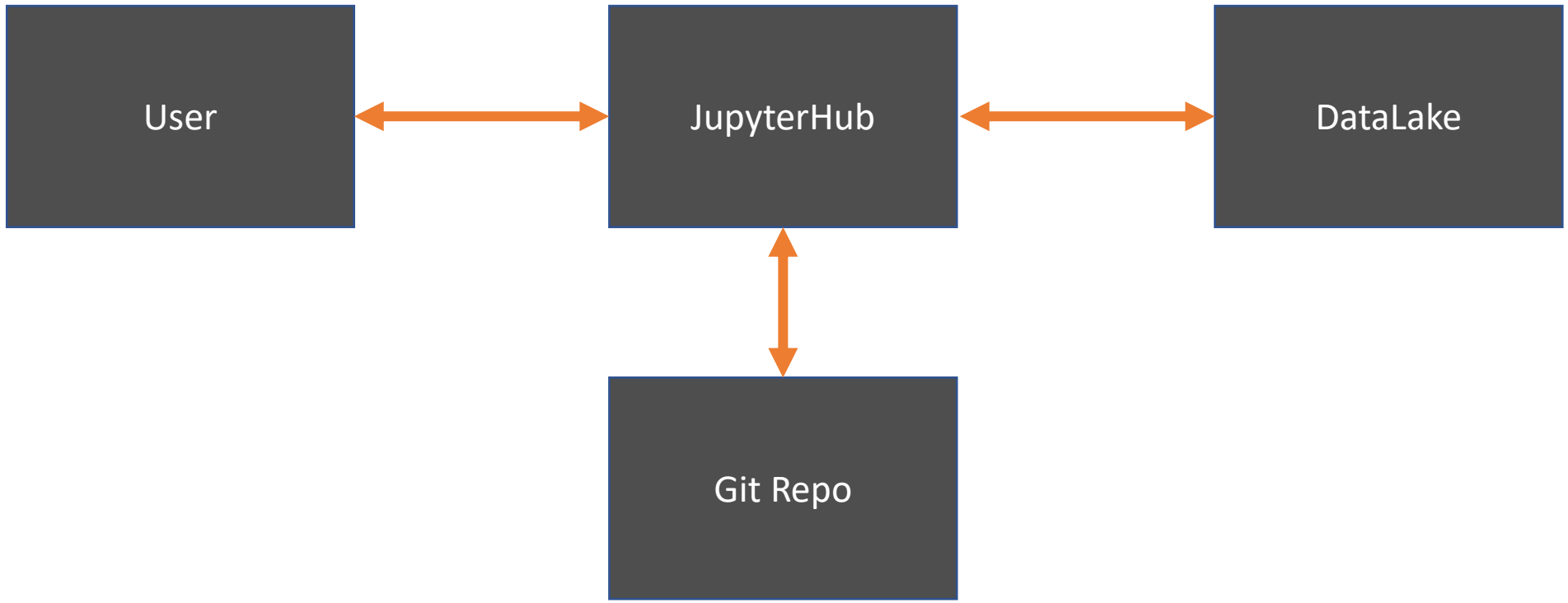

https://www.qubole.com/tech-blog/apache-airflow-tutorial-dags-tasks-operators-sensors-hooks-xcom/

# 6. Future Directions

Solution Space


2
INSTRUMENT
FILE OBJECT
SSO
1
SEARCH ENGINE
OUUID
SUUID
FUUID?
+ EXTRA METADATA
+ NOTES
LAB NOTEBOOK
OUUID
SUUID
FUUID?
+ NOTES
OUUID
SUUID
DJANGO APP
OUUID
SUUID
FUUID
FILE OBJECT
3
FOLDER WATCHER
OUUID
SUUID
FUUID
+ EXTRA METADATA
+ NOTES
4
OUUID
SUUID
FUUID
FILE OBJECT
S3
5
DATABASE
OUUID OWNER UUID
SUUID SESSION UUID
FUUID FILE UUID

# jupyterhub



# Continued Improvement

- Improved data management architecture
- Get buy-in for having more dedicated data governance roles
- Continually evolve data governance policies

# 7. Summary

## Problem Space

- Diverse requirements
- Non-conventional use case
- Overlapping roles

## Solution Space

- Common Data governance Philosophies
- Common Data Governance Architectures
- Lots of tech needed
  - Docker, Object storage (S3), RDBMS (PostgreSQL), Data catalogue (Amundsen), Orchestration (Airflow)
- Change
  - New roles, new infrastructure, new concepts for data management in academia