The PUNCH4NFDI Consortium

Overview of methods and frameworks for storage of metadata in the petabyte range

Tim Oelkers (HTW Berlin)

Hermann Heßling (HTW Berlin)

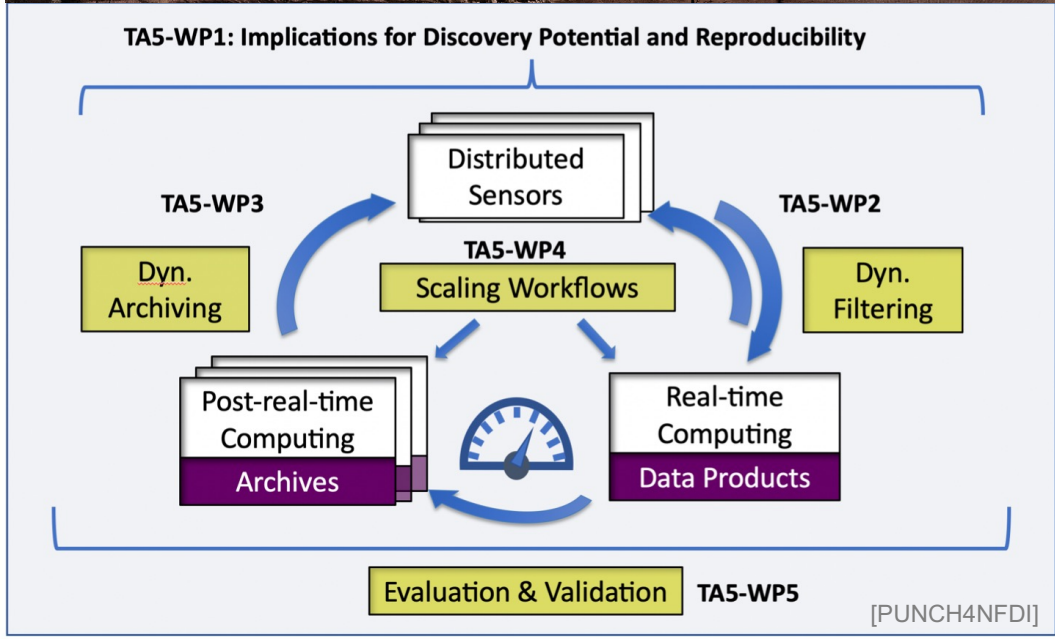
PUNCH Lunch, 30.11.2023

TA5: Data Irreversibility

- Strong reduction of data volumes during data acquisition
- Challenge: **irreversible loss of information**

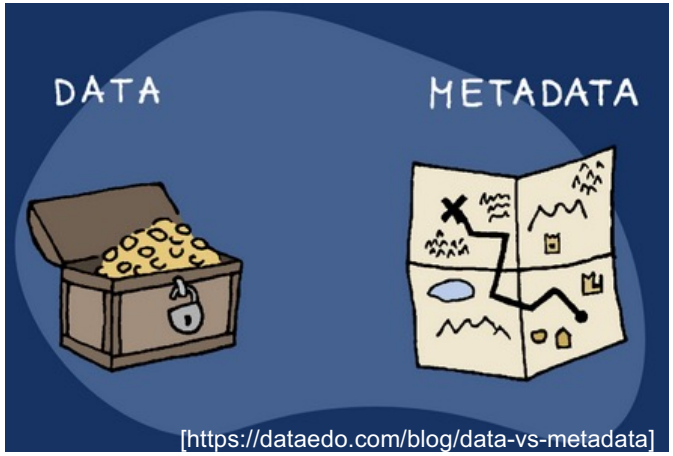
Strategies

- Dynamic Filtering
  - Extracting “interesting” information, and steering sensors in real-time
- Dynamic Archiving
  - Feedback to sensors in (near-) real-time
- Scaling Workflows: parallelization



Dynamic Life Cycle Model

Metadata Challenges



[https://dataedo.com/blog/data-vs-metadata]

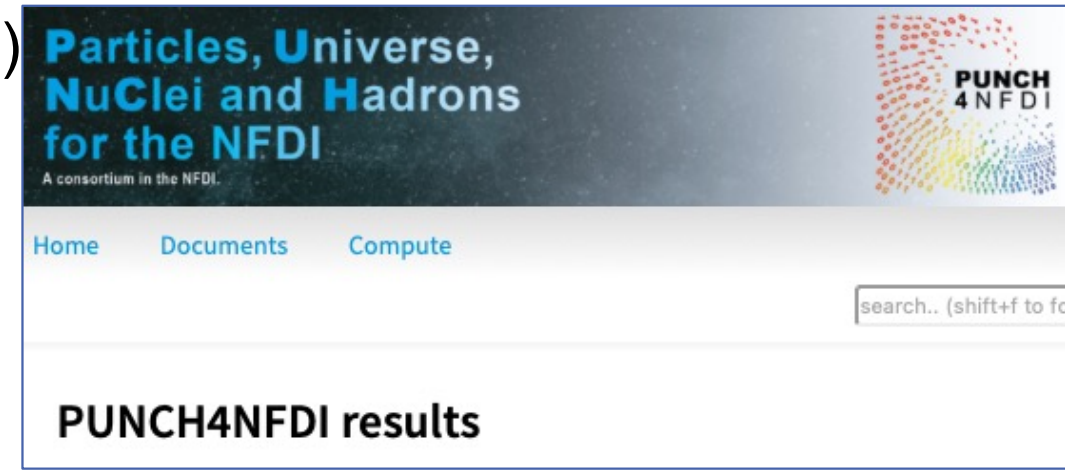
Metadata

- Traditional understanding: data about data
  - Information about experiments (+ environment)
- Extended understanding: reflect level of Data Irreversibility
  - From lost (raw) data, no information can be recovered
- Archived data are (mostly) derived data
  - Details about selection criteria and reduction pipelines important for understanding
- Strong increase of metadata volumes
  - Astronomy: boundary conditions (noise signals, weather, satellites, …) can change dynamically and unpredictably
  - Details about real-time (filter) algorithms, pipelines, “smart sensors”, …

Metadata @ PUNCH4NFDI

TA5

- Living document: *Curation and Metadata – Concepts for Data Irreversibility*
- Workshop: Jan. 18+19, 2024 (Berlin)



TA6-WP3: *Cross-community efforts towards FAIR data*

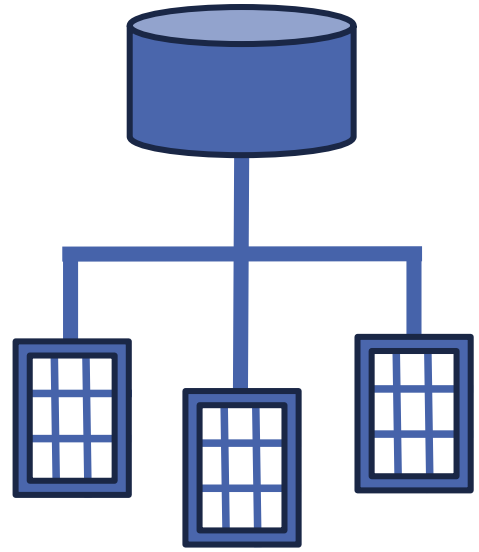
- Tim Oelkers: *Overview of petabyte-scale metadata storage methods and frameworks*

https://results.punch4nfdi.de/files/documents/Metadata_PUNCH_Oelkers.pdf

Data Requirements in Astronomy

- The High Energy Stereoscopic System (H.E.S.S.)
  - 300 GB per hour, 6 hours per night
  - 4 Mpx images, stored in filesystem with ROOT
  - Metadata in MySQL database, thousands of parameters (date, weather, parameters of telescope)
  - Refilling of database if parameters change, 2x a year

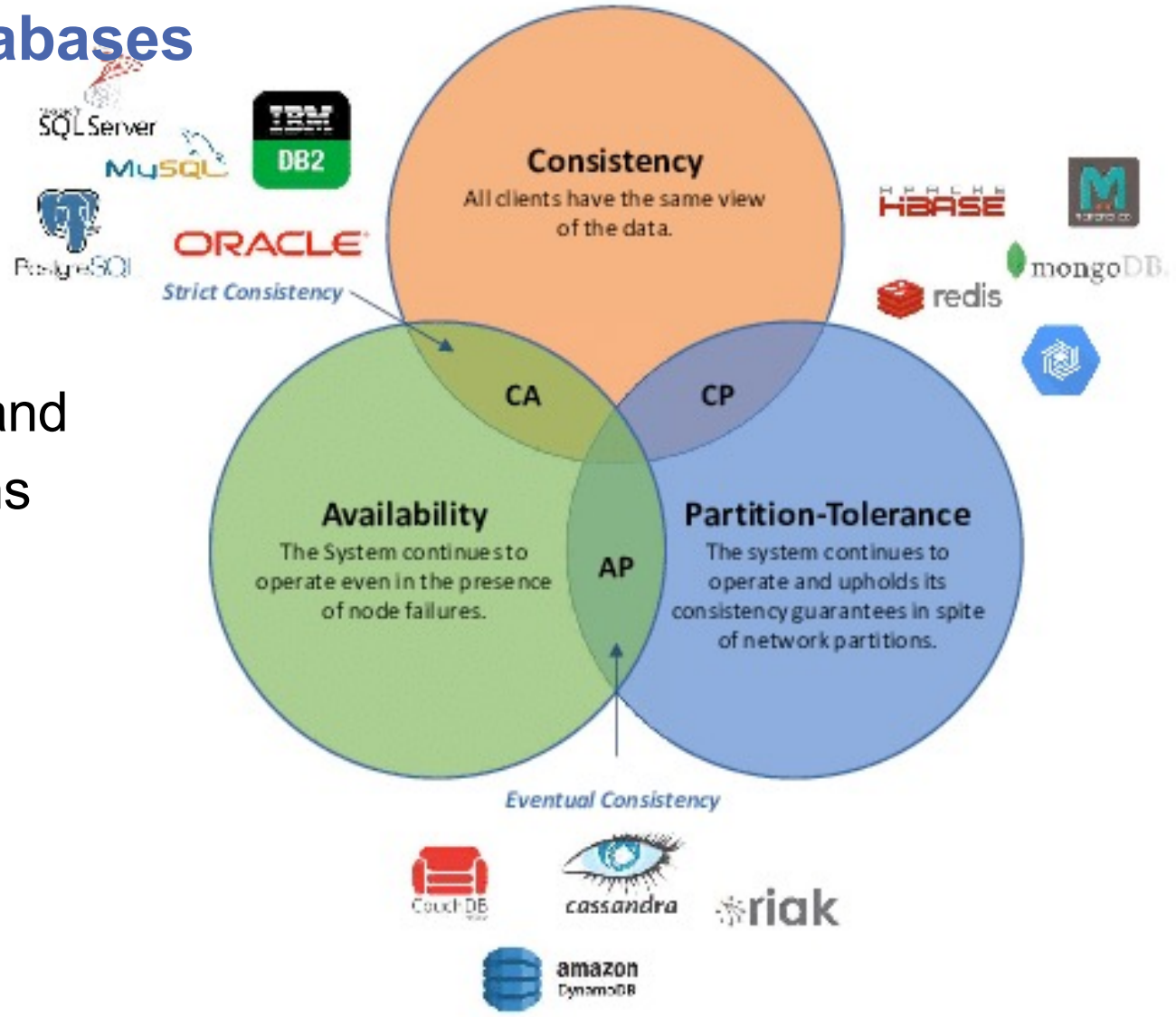
Problems with SQL Databases



- SQL as a query language requires relational, structured data
- Fixed schema
  - normalisation, modelling of relations in advance
  - inflexible to changes in requirements
- Weak scalability
  - usually limited to 1 physical server
  - longer access times with increasing data volumes and number of users

Problems with SQL Databases

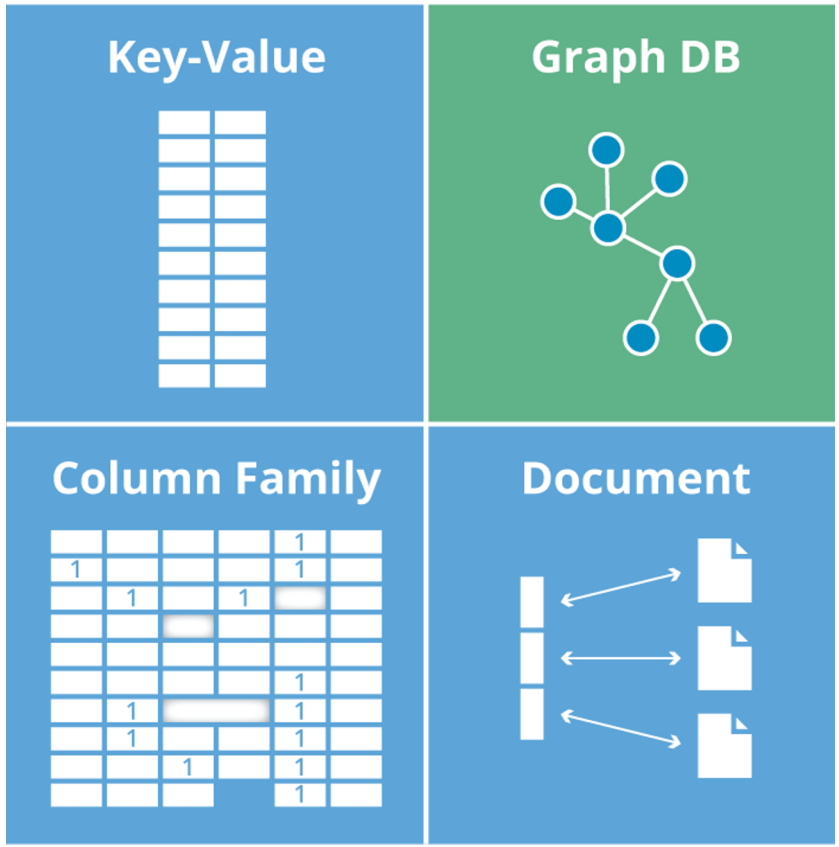
- ACID: Atomicity, Consistency, Isolation and Durability ⇒ transactions
- CAP theorem
- RDBMS focus on strict consistency by design



[https://www.debadityachakravorty.com/system_design/captheorem]

NoSQL Databases

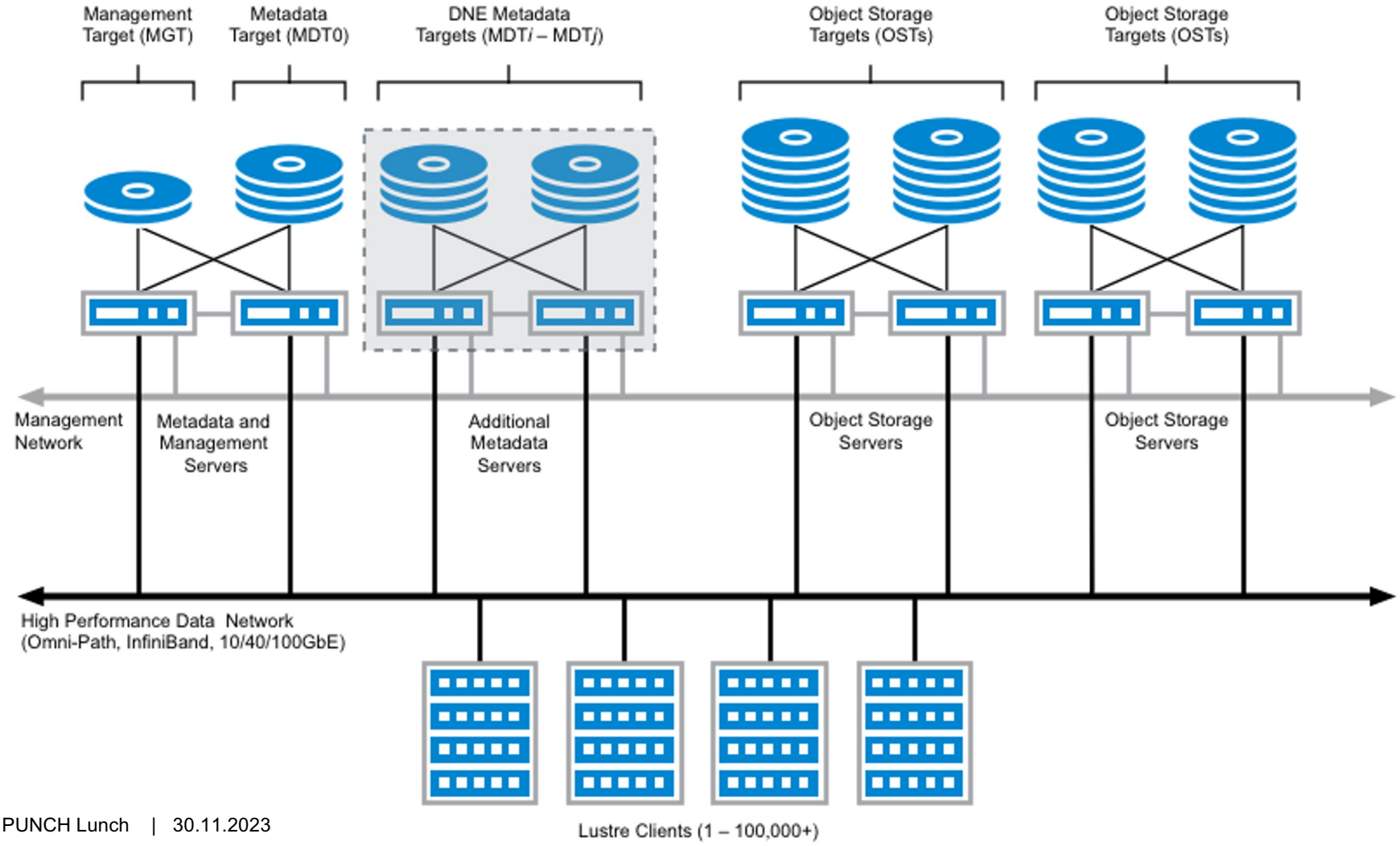
- NoSQL allows different data structures
- BASE: Basically Available, Soft state, Eventual consistency ⇒ no transactions
- No normalisations: relations are modelled as separate entries



Distributed Filesystems: Lustre

- Filesystems: no queries, no DB overhead ⇒ archiving
- Direct access to files ⇒ better integration in other frameworks
- Lustre: used by 60 out of Top100 supercomputers
- Transparent distribution of data with POSIX compatibility
- File striping, cache coherent, multi-layered storage, fault tolerance

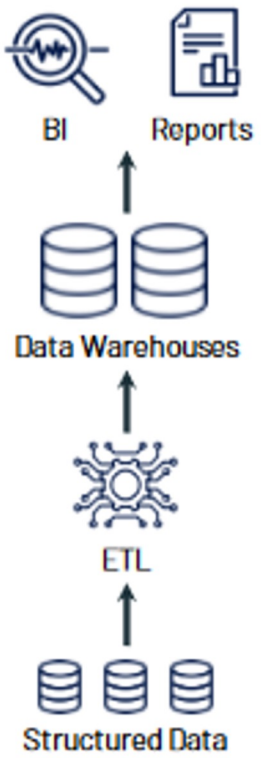
Distributed Filesystems: Lustre



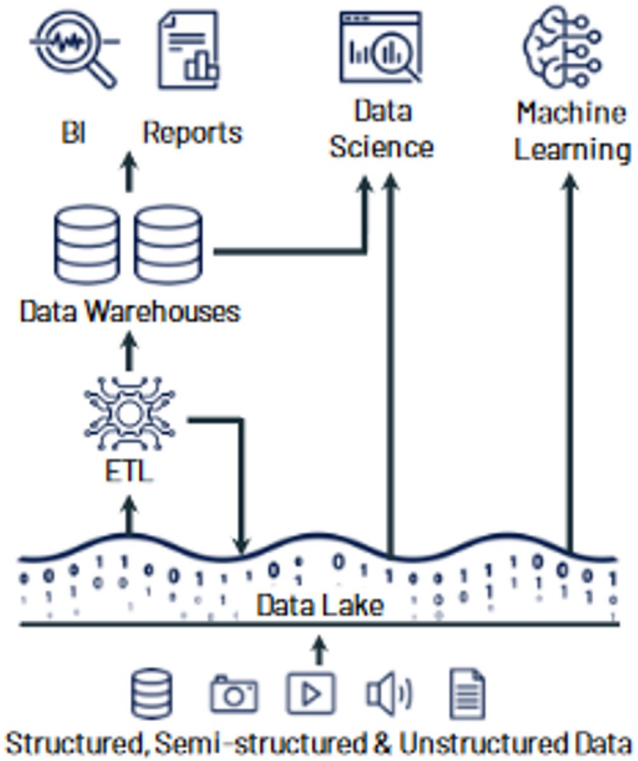
File format: Apache Parquet

- Open Source, column-oriented file format ⇒ excellent read times and compression
- Excellent integration into Big Data frameworks like Apache Spark, Pandas
- Split of metadata and data ⇒ data can be distributed to different files
- Apache Thrift for definition of data types

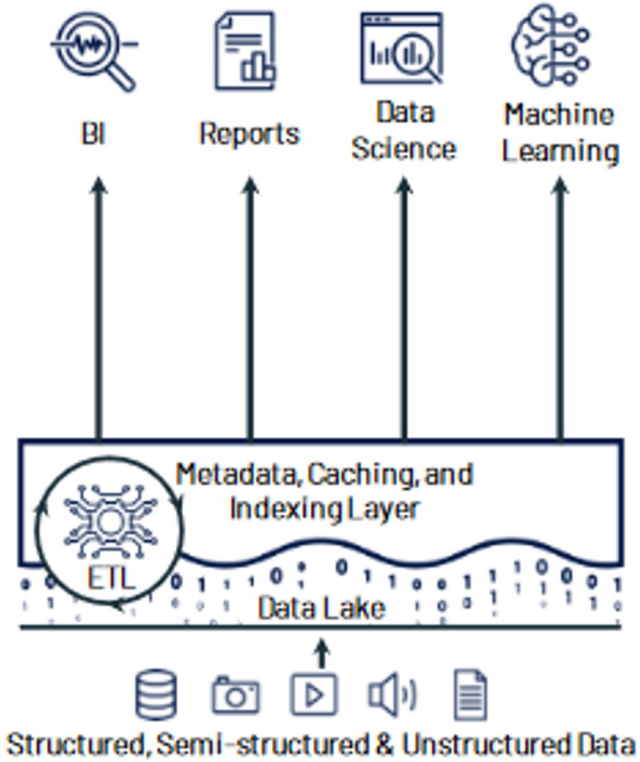
Lakehouse: Delta Lake



(a) First-generation platforms.



(b) Current two-tier architectures.



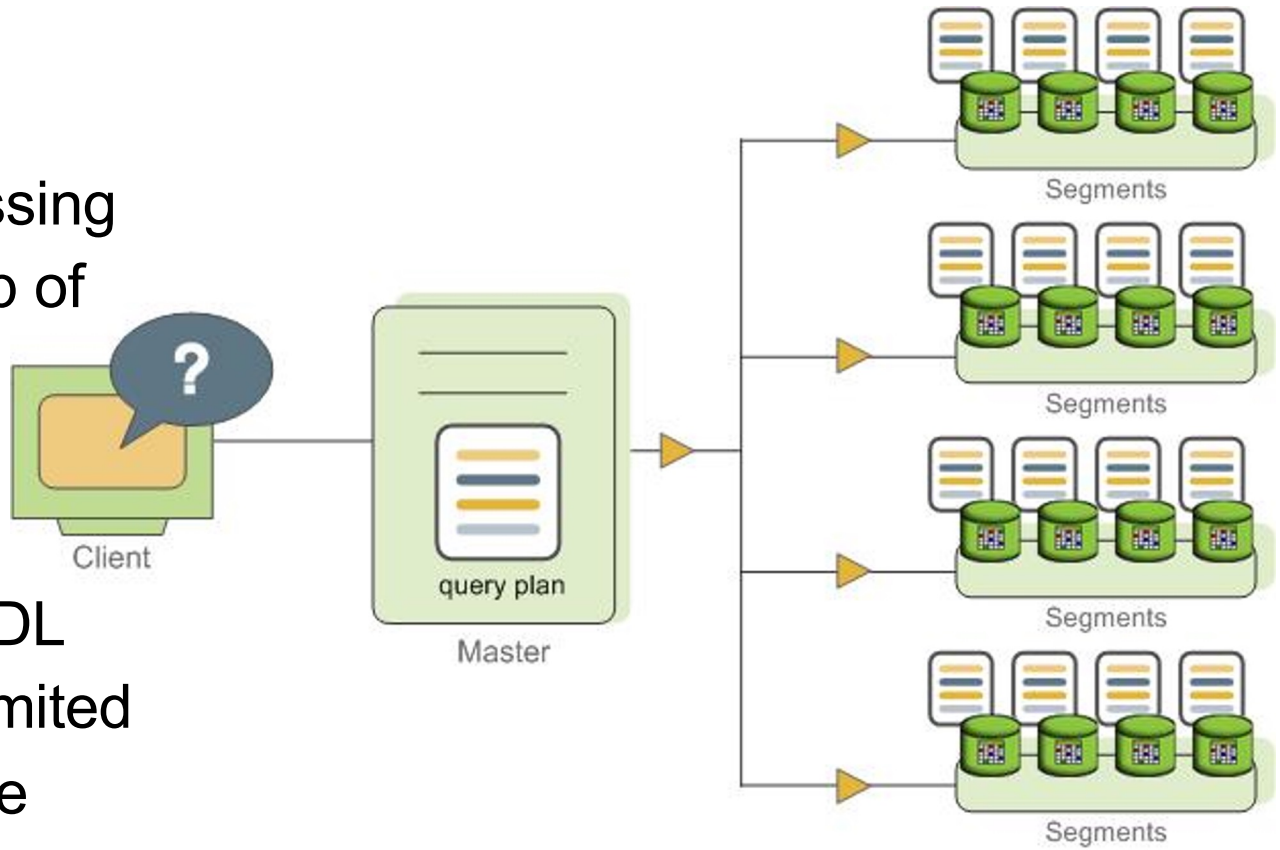
(c) Lakehouse platforms.

Lakehouse: Delta Lake

- Direct access to data lakes in open-access formats
- ACID compatibility, versioning, indexing, machine learning, caching
- Medallion Architecture: pipeline that incrementally restricts schema
  ⇒ gold layer contains aggregated & joined & filtered data ready for specific use case
- Faster and more performant than AWS, Azure, Google Cloud

Federated databases: Greenplum

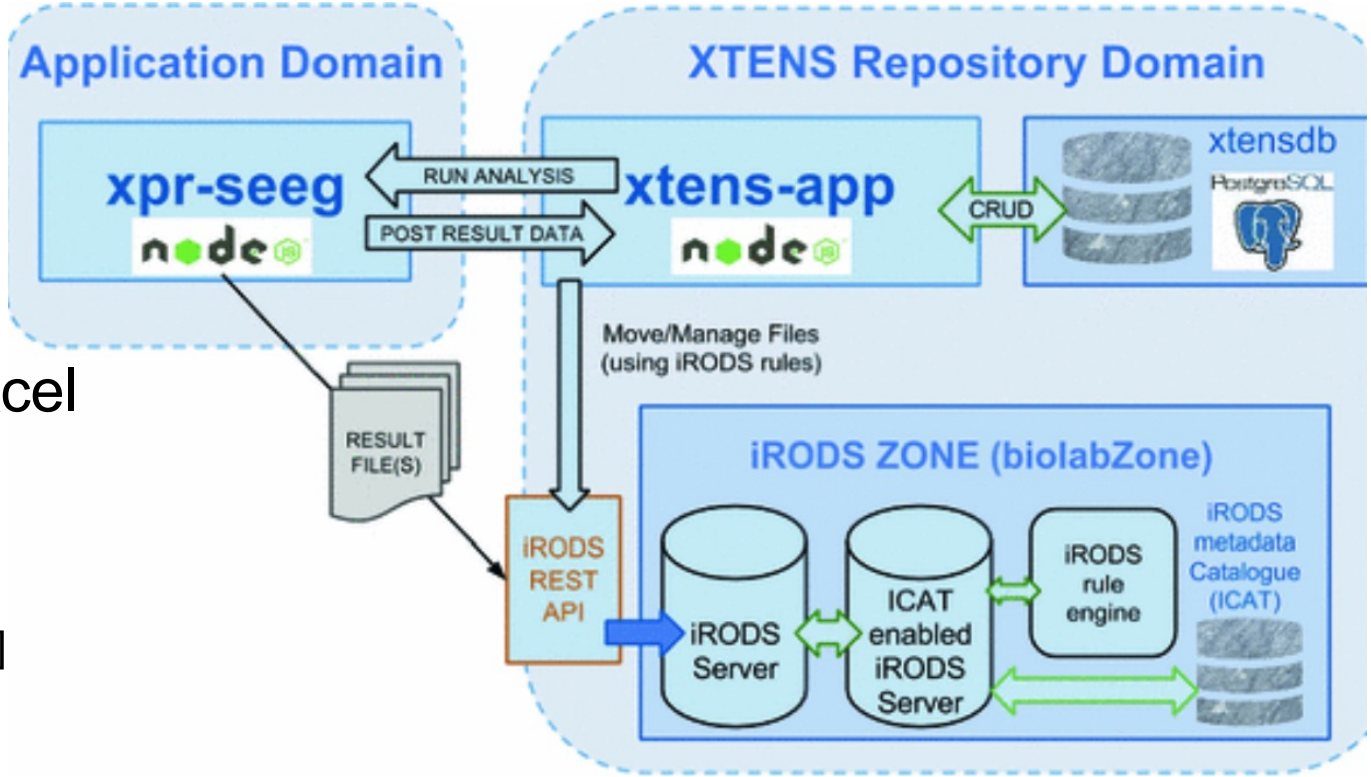
- Massively parallel processing architecture (MPP) on top of PostgreSQL
- Query planner GPORCA designed for Big Data
- Distribution defined by DDL
- ACID compatibility, but limited transactional performance



Metadata : specific frameworks

XTENS 2

- Repository for biomedical data
  - Managing heterogenous data (samples, any kind of data)
  - Complex queries
- Database: PostgreSQL
  - **Flexible JSON data model**
  - XTENS 1: MySQL, Excel
- Distributed data: iRODS



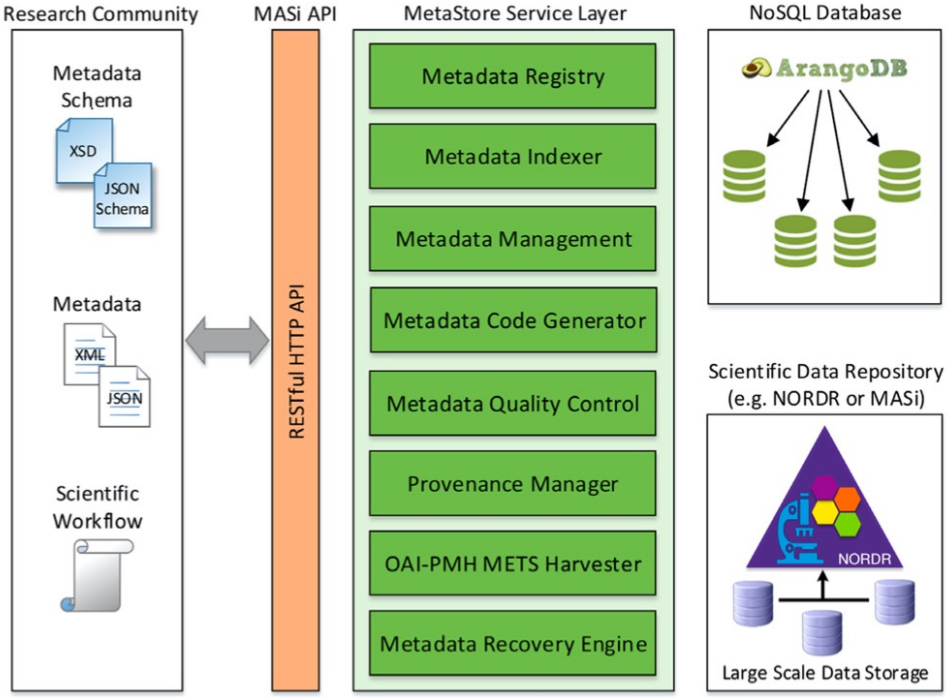
[https://github.com/xtens-suite/xtens-app]

Metadata : specific frameworks

MASi (Metadata Management for Applied Sciences)

- Generic metadata programming interface
- Open Archive Initiative (OAI) protocol for metadata harvesting
- Multi-community support: geography, chemistry, digital humanities
- Multi-model NoSQL DB
  - Key-values
  - Documents
  - Graphs
- Future (2019): integration into UNICORE (⇒ interface to HPC)

[https://tu-dresden.de/zih/forschung/projekte/masi]



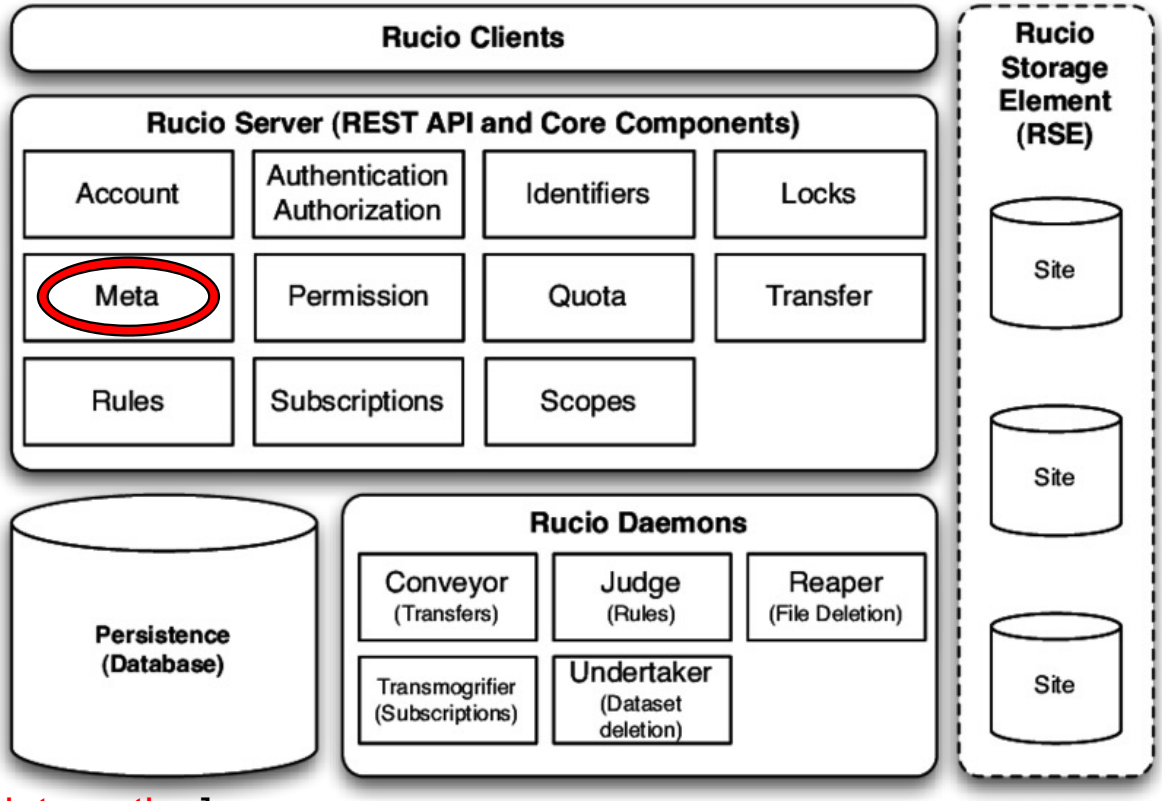
Metadata : generic framework

Rucio (LHC@CERN)

- Highly scalable data management framework
- Special Interest Group (SIG) on metadata (2021-2023)
  ⇒ Publishing Rucio metadata as an IVOA service

IVOA (International Virtual Observatory Alliance)

⇒ standards for metadata



[https://github.com/rucio/rucio]
[https://rucio.cern.ch/documentation/sig_metadata]
[https://gitlab.com/ska-telescope/src/ska-rucio-ivoa-integration]

Summary

- Personal evaluation of H.E.S.S.
  - Filesystem great for long term storage ⇒ Lustre: easily integrable, better performance
  - Apache Parquet instead of ROOT for FAIR data access
  - NoSQL (document store) instead of MySQL
  - Greenplum useful for scaling relational databases
- Specialized metadata frameworks: XTENS 2, MASi
  - Little / no further development
- Generic metadata framework: Rucio
  - Activities to integrate metadata requests from other communities
  - Integration of "dynamic metadata" in the petabyte range feasible?

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)

Project number: 460248186 (PUNCH4NFDI).

Thank you!

The PUNCH4NFDI Consortium

Spokesperson:
PD Dr. Thomas Schörner (DESY, thomas.schoerner@desy.de)
DESY, Notkestr. 85, D-22607 Hamburg

Contact:
Mail: punch4nfdi@desy.de
Web: www.punch4nfdi.de
Twitter: #punch4nfdi