

BACKGROUND & CHALLENGE

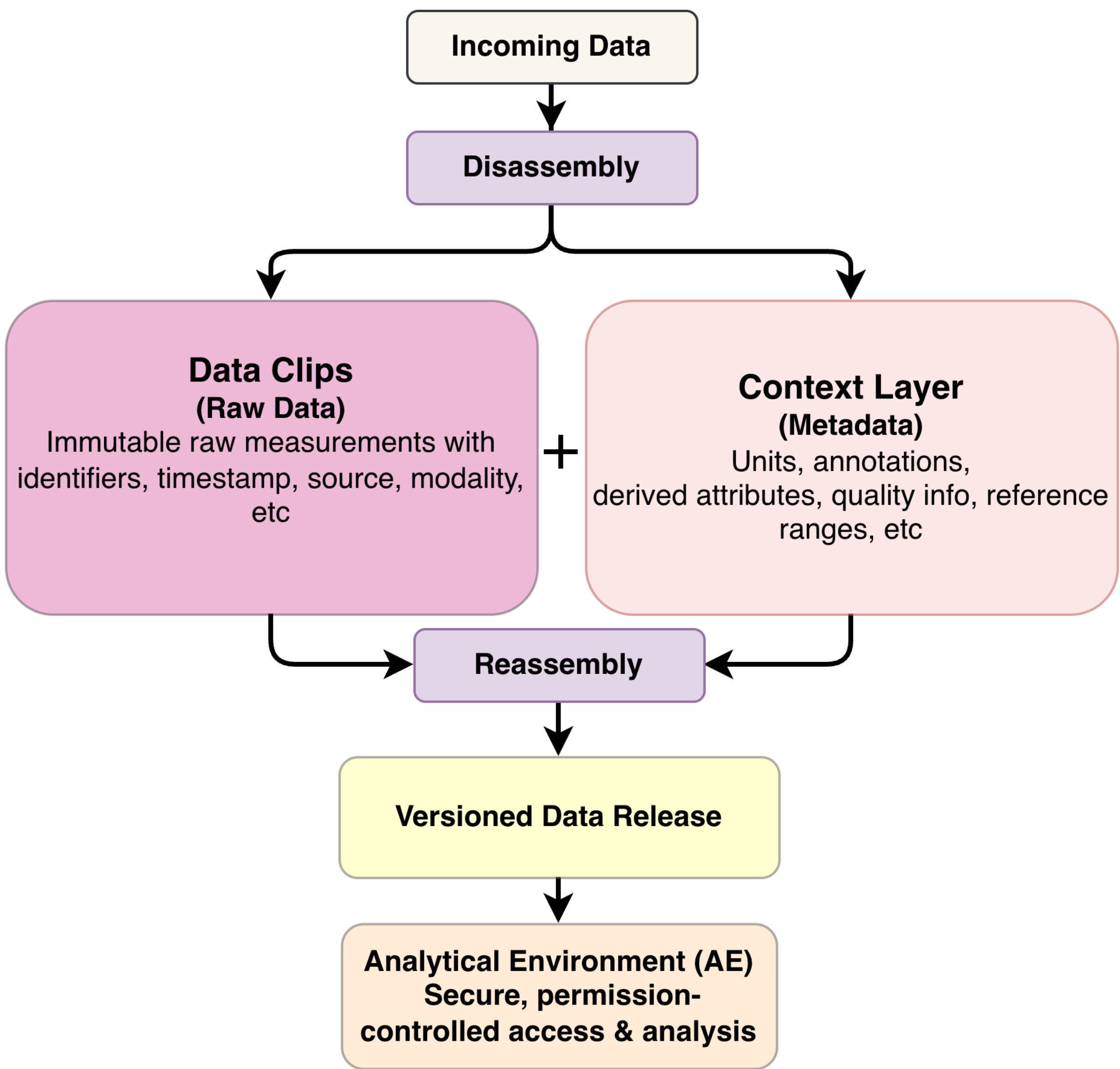
Why is this needed?

- Asynchronous data submissions
- Multiple data providers & devices
- Continuous corrections & updates
- GDPR requirements
- Reproducibility challenges
- Provenance tracking difficulties

Problems with Traditional Pipelines

- Dataset duplication
- Version inconsistencies
- Poor accountability
- Difficult synchronization

FRAMEWORK OVERVIEW



Design Principles

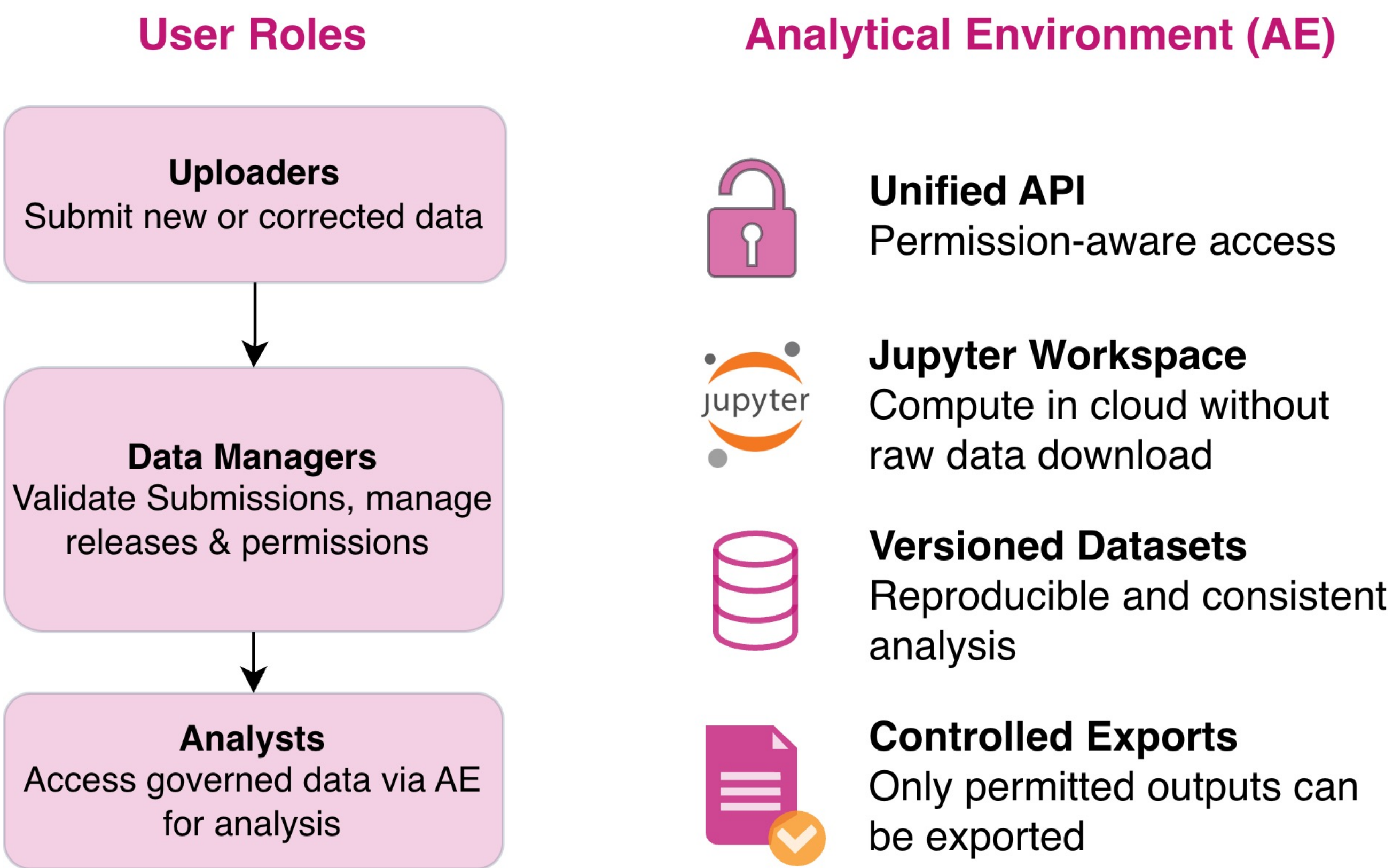
- Immutable Measurements**
 - Raw observations are preserved and never overwritten.
- Context Separation**
 - Interpretation and metadata evolve independently of raw data.
- Selective Updates**
 - Only affected clips or context entries are modified.
- Full Provenance**
 - Every change is logged, versioned and completely traceable.

Standard Compatibility

Clips can be reassembled into standard formats, such as CDISC and FHIR.

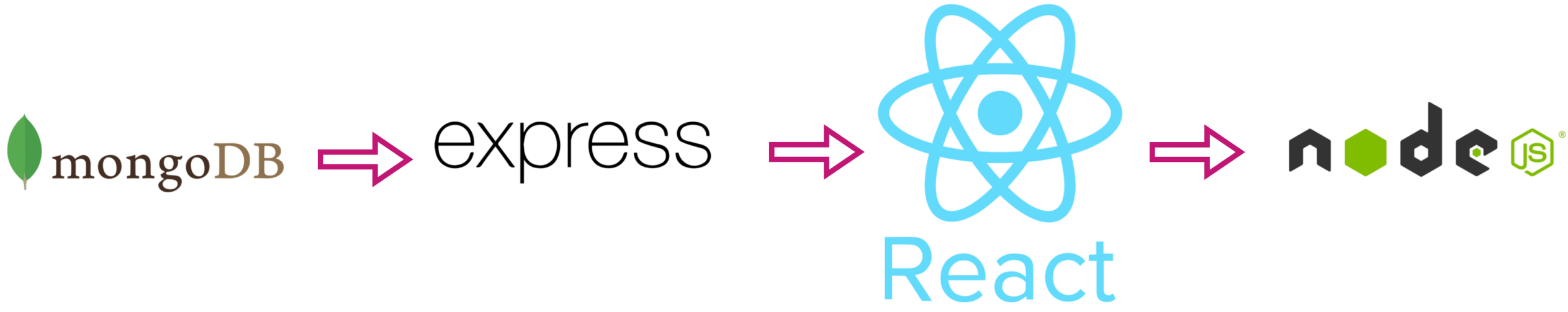


GOVERNANCE & ANALYSIS



IDEA-FAST DEPLOYMENT

IDEA-FAST DMP Implementation uses the MERN architecture: MongoDB, Express, React, Node.js



A cloud-hosted centralized platform that has been operational since 2020
Deployment at Scale

- As of Oct 2025, the DMP has
- Managed more than 3 million data clips
 - Taken care of the data of more than 1900 patients
 - Covered data from more than 60 Organisations
 - Authenticated more than 350 users
 - Made 81 weekly governed release

Operational Efficiency

Despite continuous updates from 60+ organizations and millions of records, the entire governance workflow (including ingestion checks, release management, permissions and support) can be managed by 1-2 data managers with full provenance.

Security & Resilience

- Independent Hot Backup**
 - Externally operated backup providing autonomous recovery path
- Air-Gapped Tape Backup**
 - NDMP tape library backup; full write/read cycle -21 days
- Multi-Region Mirroring**
 - Read-only replication & snapshots for rapid rollback and protection

DEVICE FEATURE EXTRACTION PROGRESS

Feature	Device	Description	Progress
HRV	VitalPatch	Derived HRV features from raw ECG signal obtained from VitalPatch	100%
VIT	VitalPatch	Basic derivatives from VitalPatch HR, Temp, RR, Steps	100%
COG	CANTAB APP	Derived cognitive performance metrics from the CANTAB app	100%
WKT	WildKey Keyboard	Derived typing features from WildKey Keyboard	100%
WKS	WildKey Social	Derived typing features from WildKey Social	100%
SLP	eBedSensor	Derived sleep features from eBedSensor	93%
MCR	AX6	McRoberts Raw activity classification data	92%
DMO	AX6	Derived mobility features with mobilise-d algorithm	89%
STS	AX6	Derived sit-to-stand/stand-to-sit kinematic features from the Axivity AX6 inertial sensor	87%

CONCLUTIONS & FUTURE WORK

Key Contributions

- Active governance for continuously evolving datasets
- Fine-grained clip-based architecture enables flexible and selective updates
- Version-controlled releases ensure reproducibility
- Secure analytical environment prevents raw data distribution
- Full provenance and accountability across the data lifecycle
- Proven deployment at scale across 60+ organizations with low operational

overhead

Future Work

- External deployments beyond IDEA-FAST
- Open-source adoption and community growth
- Hybrid / federated governance models
- Distributed provenance validation and replication