STARFISH

take the “meh” out of metadata

harness the “chi” in archiving

take the “rage” out of storage

put the “tada” in metadata

What Does Starfish Do?

- Starfish associates metadata with files and directories in POSIX-style file systems and in S3-style object stores.
- Metadata are used for:
  - Content classification
  - Reporting and analytics
  - User portal
  - Shaping policies and directing batch operations
- There are two forms of metadata
  - Simple tags (yes, directory tags inherit)
  - Key-value pairs (using JSON)
- Starfish sits outside of the data path so it must asynchronously maintain the inventory of the file system.
  - It works with all storage systems
  - It does not introduce any latency or points of failure
  - It is very fast and efficient
    - Starfish is suitable for extreme environments with billions of files and 100s of petabytes.

Two Main Use Cases Among Librarians and Curators

1. A middleware for storage housekeeping in a data curation facility.
   - Data protection
   - High-level content classification
   - Fixity
   - Reporting, cost-accounting, charge-back
2. A unique solution for associating metadata and automating workflows with live, mutable file storage systems.
   - Allows the curation workflow to begin with data creation
   - Facilitates the hand-off between content creation and curation
   - Facilitates the re-use of archival data sets

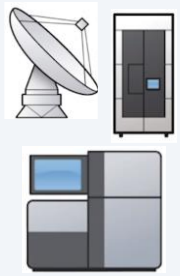
- Starfish is also great for managing files that are awaiting formal curation workflows.

Life Without Starfish

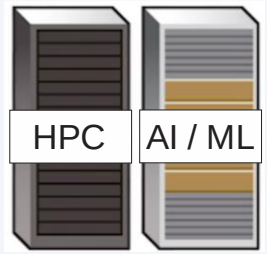
Applications that Reference Files

Content Repositories
Digital Asset Management
Data Commons
Lab Notebooks

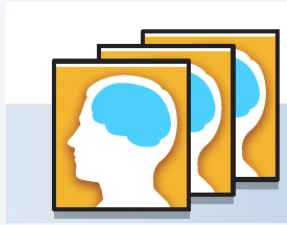
Data Acquisition Workflows



Data Processing Workflows



Tribal Knowledge (or lack thereof)

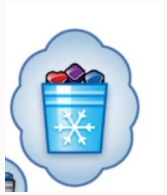


HPC Storage

Enterprise NAS

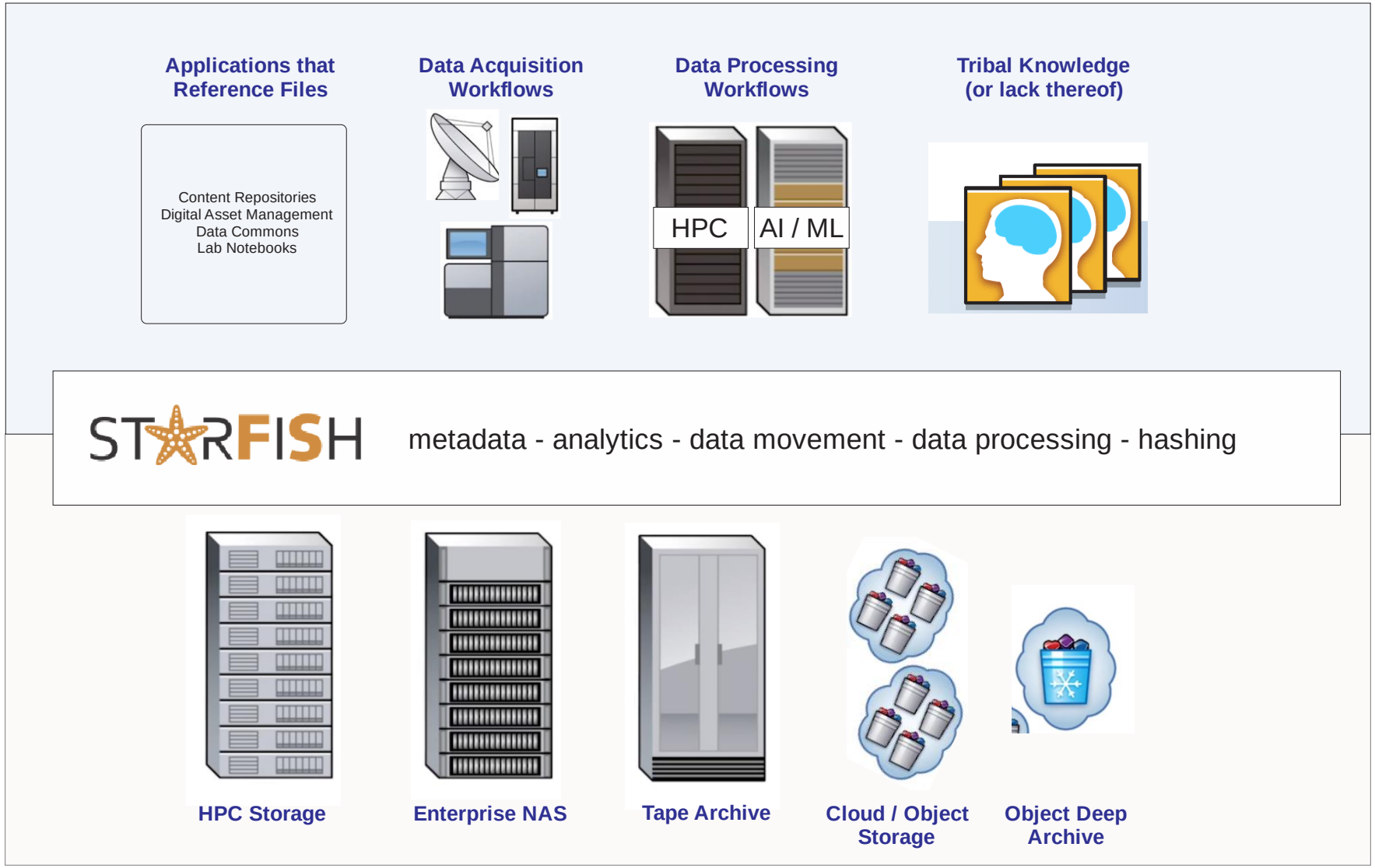
Tape Archive

Cloud / Object Storage

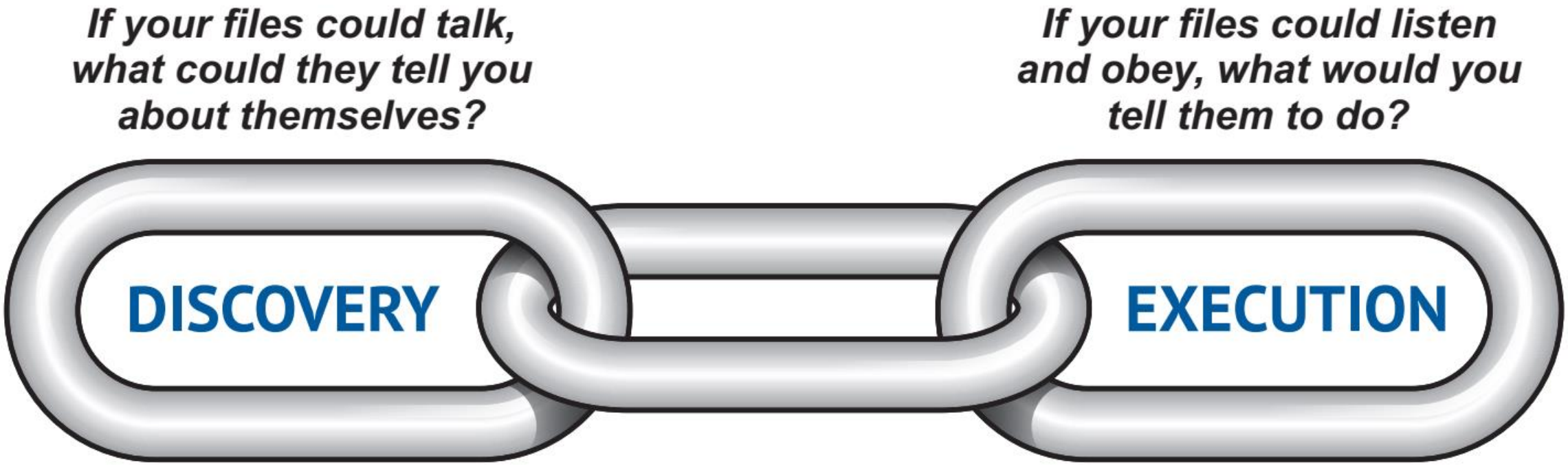


Object Deep Archive

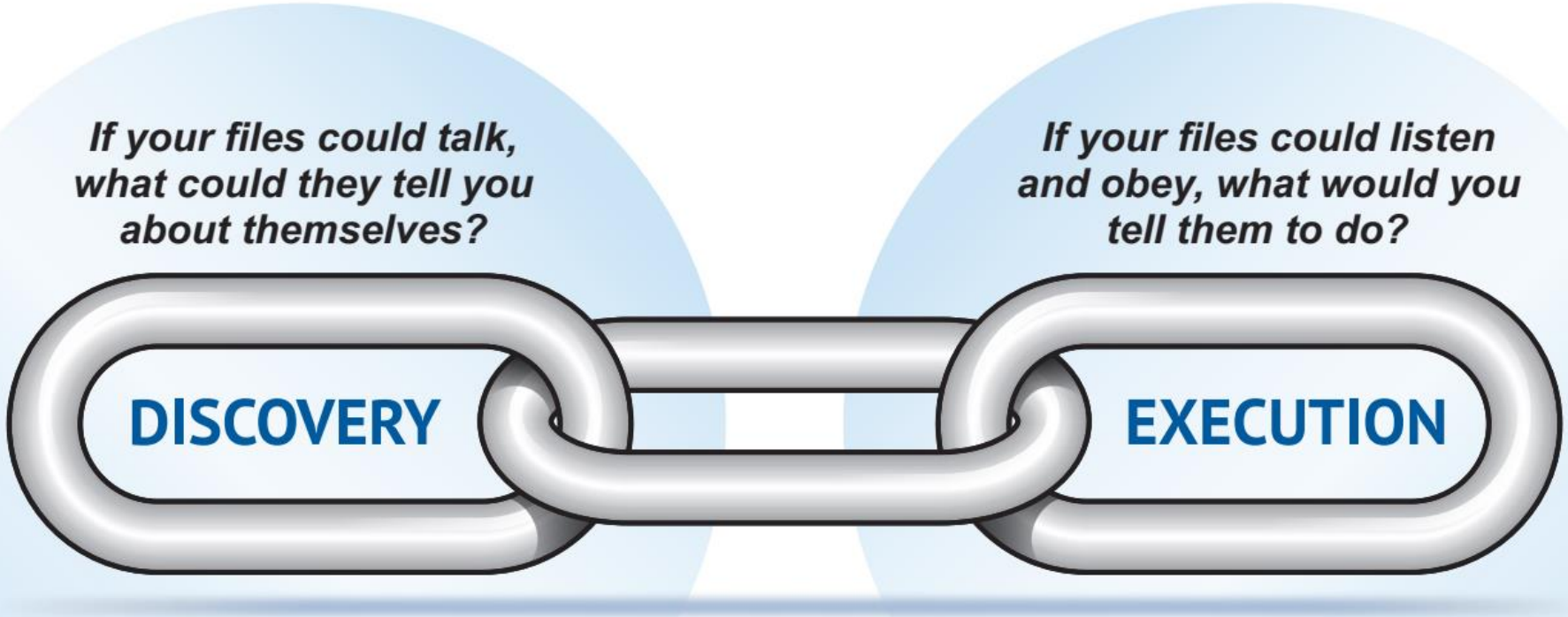
Starfish is as Storage Middleware



A Simpler Way to Explain the Technology



Discovery + Execution: A Simple But Powerful Paradigm



a data catalog for unstructured data

massively scalable - billions of files and objects
extensible metadata - tags and key-value pairs
simple and turnkey but suitable for custom integration

a scale-out data mover & batch processor

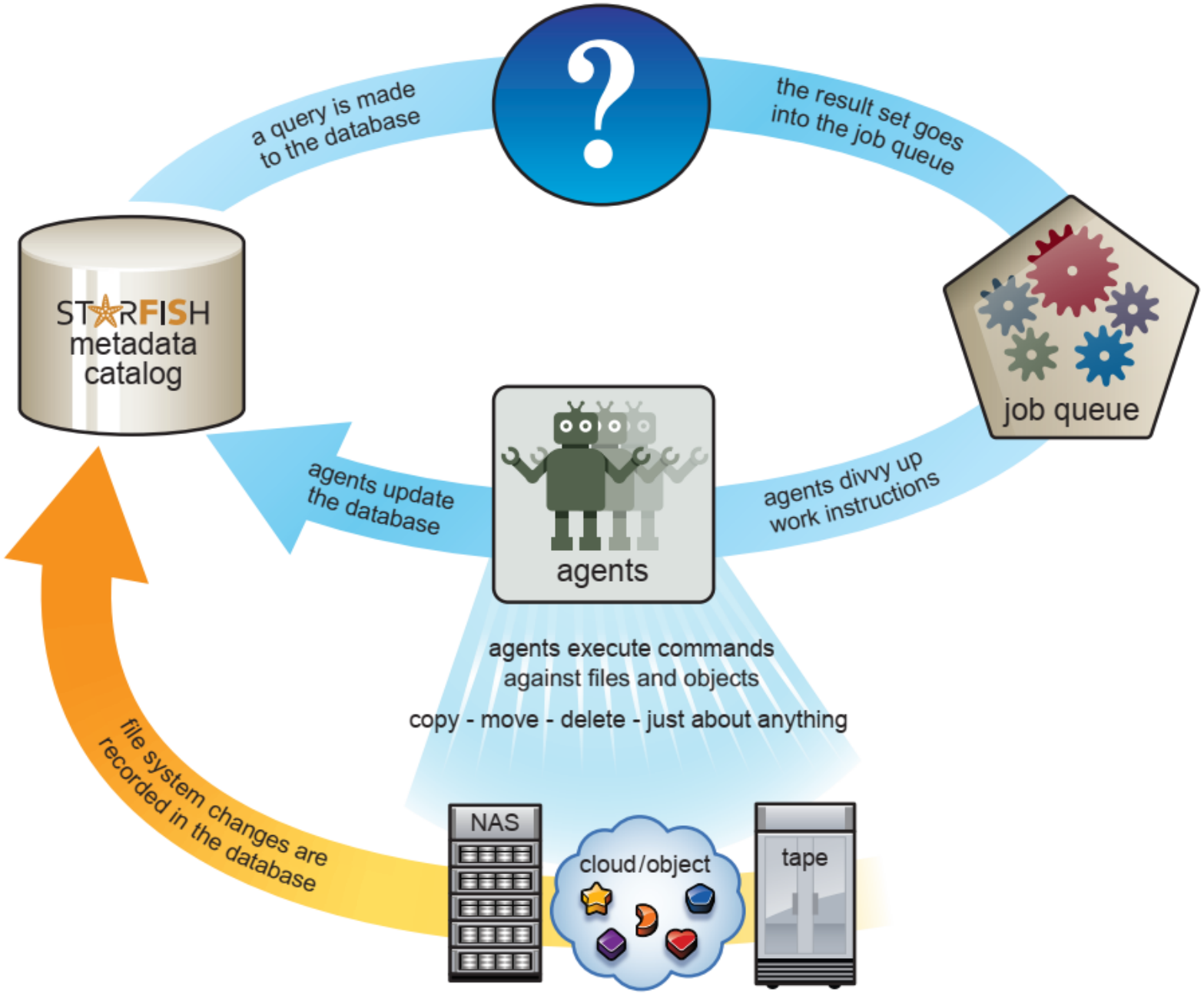
LINUX or Windows – File and Object
migrate - replicate - workflow - archive - backup - purge
easy to customize - runs your code or ours

Lather, Rinse, Repeat

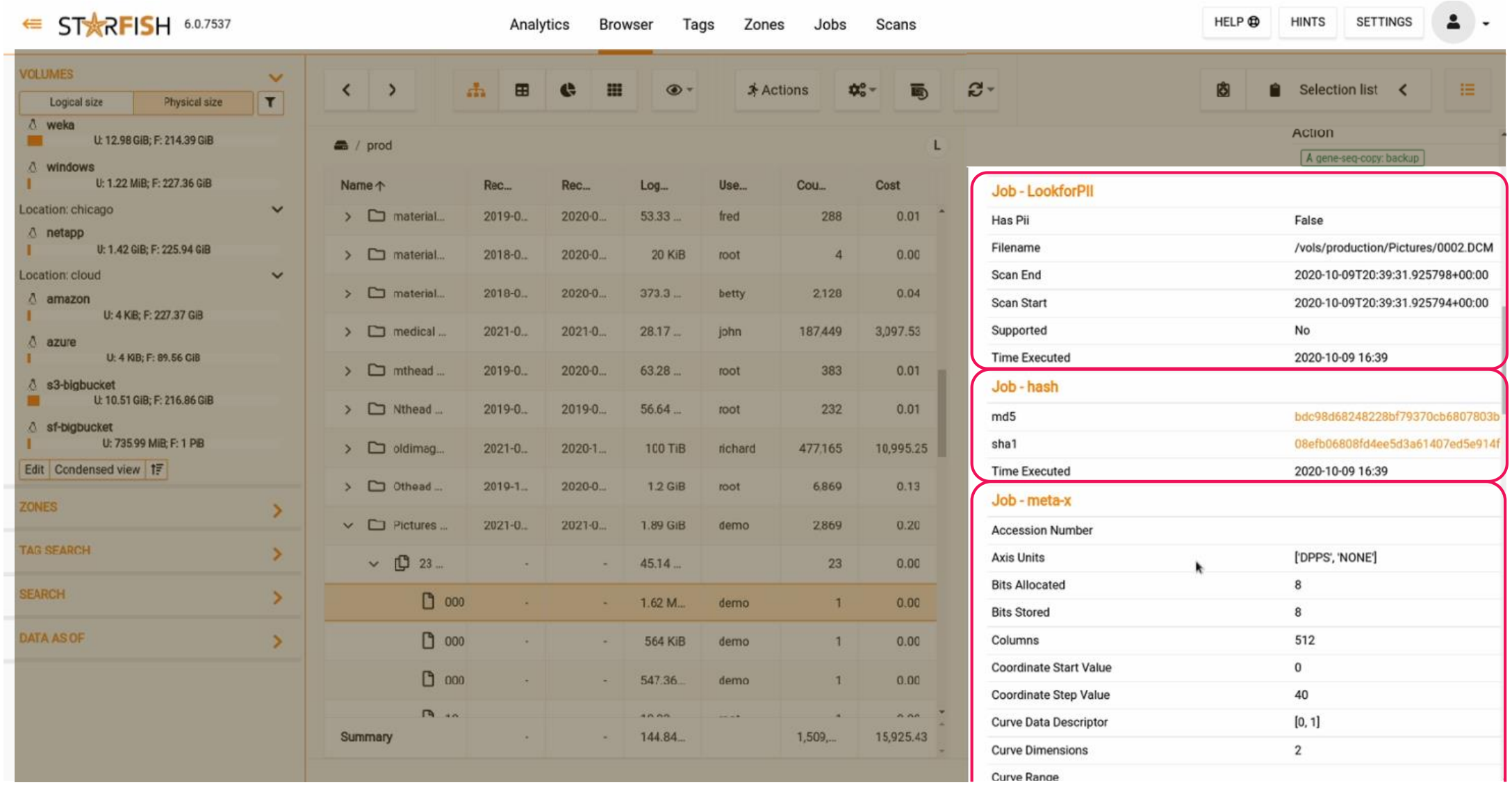
There is a feedback loop between the catalog and the jobs engine.

- Query database
- Do some stuff
- Update database

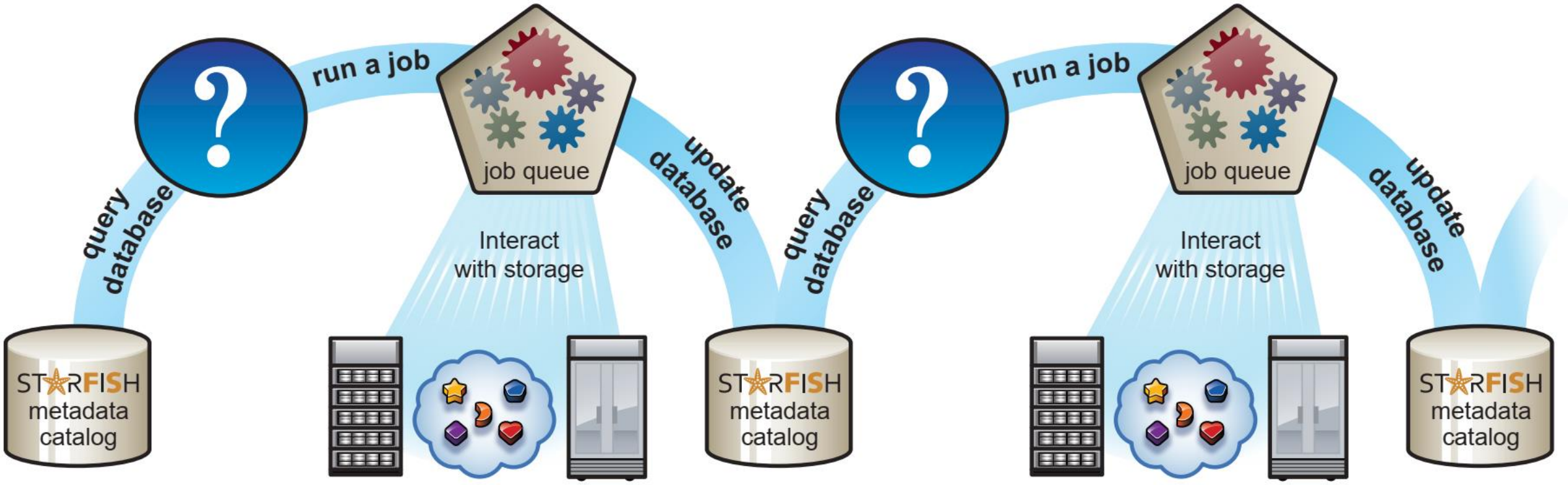
- Query the database
- Do more stuff
- Update database



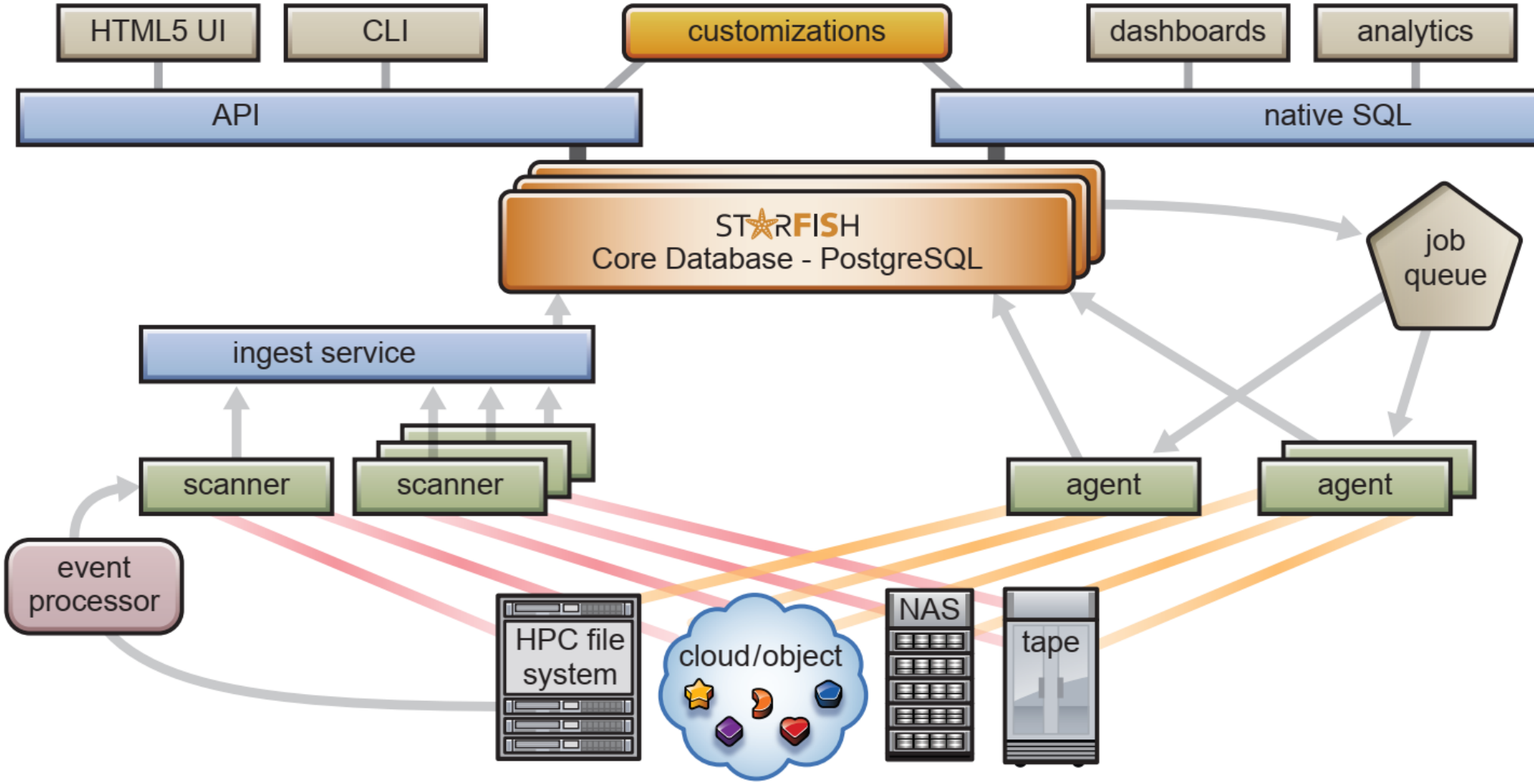
Key-value metadata is simple JSON



Knowledge Begets Knowledge



Starfish Topology: Designed for Scale and Customization

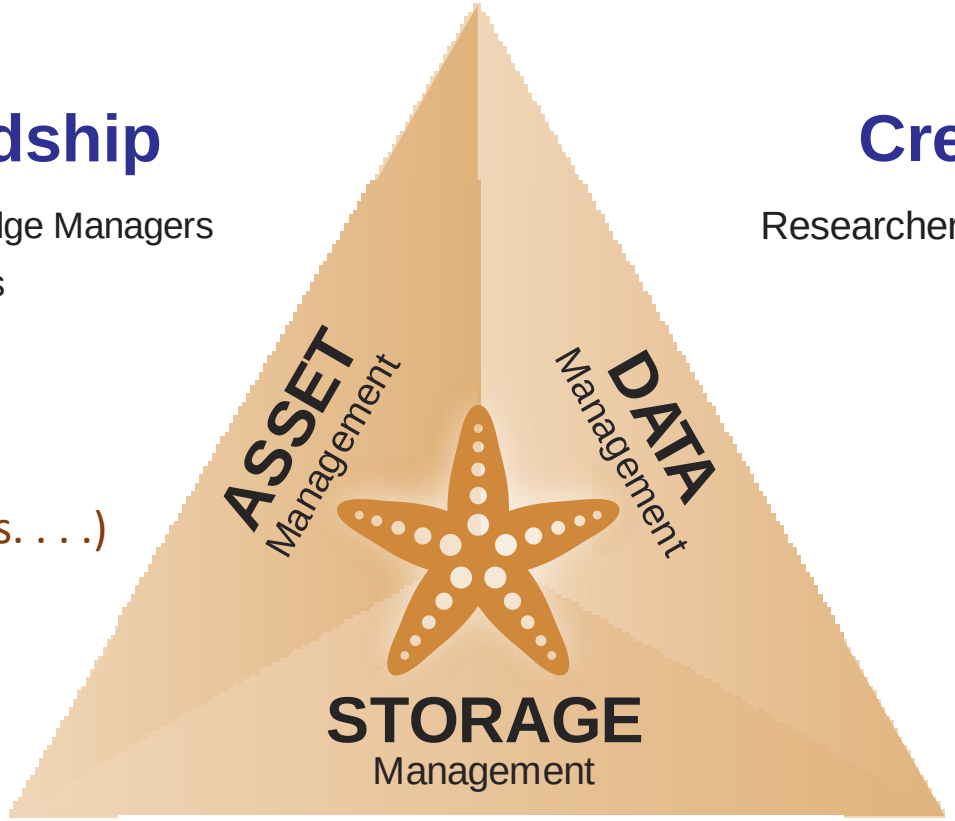


The Grand Vision

Governance & Stewardship

Archivists - Curators - Librarians - Knowledge Managers
Data Governors - Auditors - Data Wranglers

Retention / Disposition
Anonymization / Privacy
Compliance (GDPR - Export Controls. . . .)
Data integrity assurance
Content classification
Data provenance
Data cataloging



Creation and Consumption

Researchers - Artists - Animators - Engineers - Quants
Core Facilities - Data Scientists

Metadata tagging
Search
Workflow automation
Data provenance
Collaboration
Data reusability
Data transmission

Sys Admins (Storage, Backup, HPC) - IT Directors

Data Movement: Archiving/Tiering, Migration, Replication, Cloud-bursting, Backup & Restore

Reporting: Capacity Planning, Aging, Cost Accounting, Charge-back / Show-back

Misc Operations: Permissions Management, GDPR Compliance, Job Scheduling

Starfish is Open, Low Risk and Built for Science and R&D

- Non-proprietary
  - No proprietary file formats
    - Industry standard compression (GZIP and LZ)
    - TAR when group files together
  - Storage device agnostic
  - Metadata is stored in SQL (Postgres)
- Starfish is out of the data path
  - It does not introduce latency
  - It does not introduce points of failure
- Starfish was built for Research Data Management
  - Our long term roadmap addresses the coming challenges of research computing