take the “meh” out of metadata

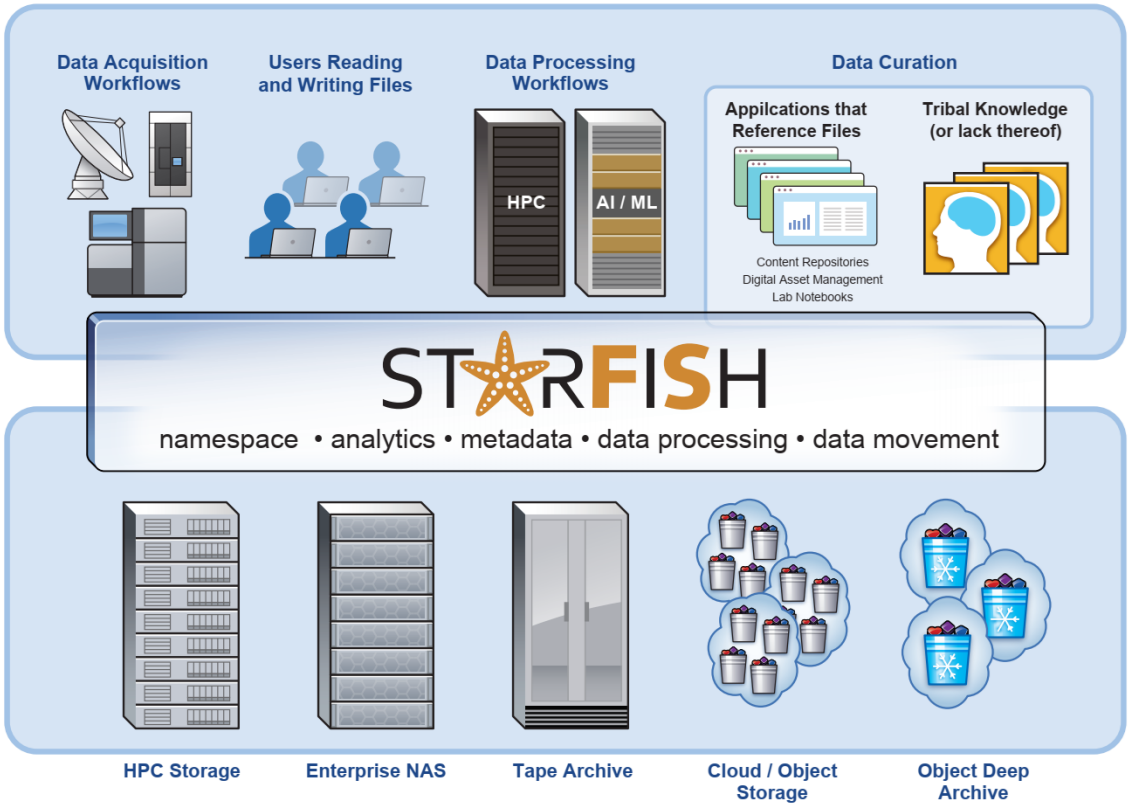
harness the “chi” in archiving

take the “rage” out of storage

put the “tada” in metadata

www.StarfishStorage.com

Starfish Logically Federates The Storage Environment



Starfish enables the entire storage environment to work logically together.

Starfish moves files between devices without adding to the entropy.

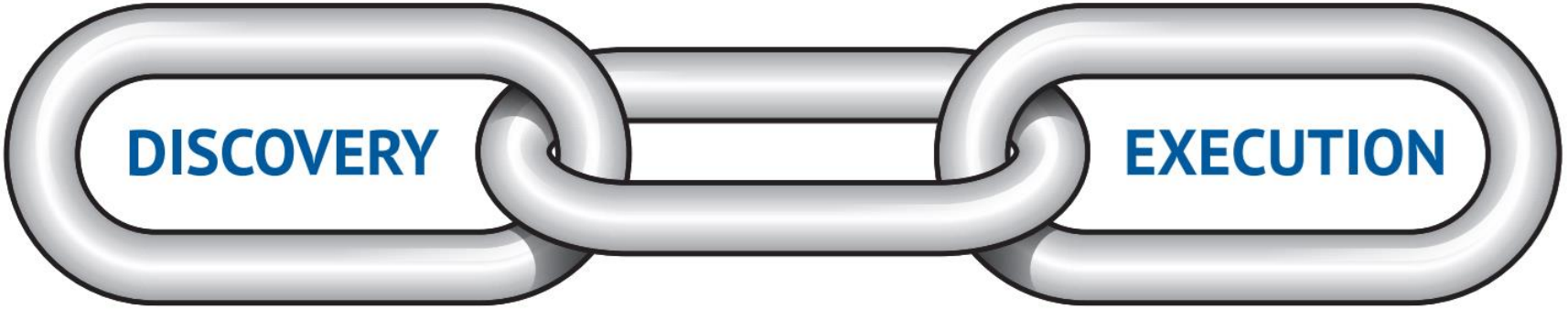
Data curation begins when files are first created.

Unnecessary files and excess copies are deleted, while backups and gold copies are preserved on the appropriate media.

Starfish: A Simple, but Powerful and Versatile Paradigm

If your files could talk, what could they tell you about themselves?

If your files could listen and obey, what would you tell them to do?

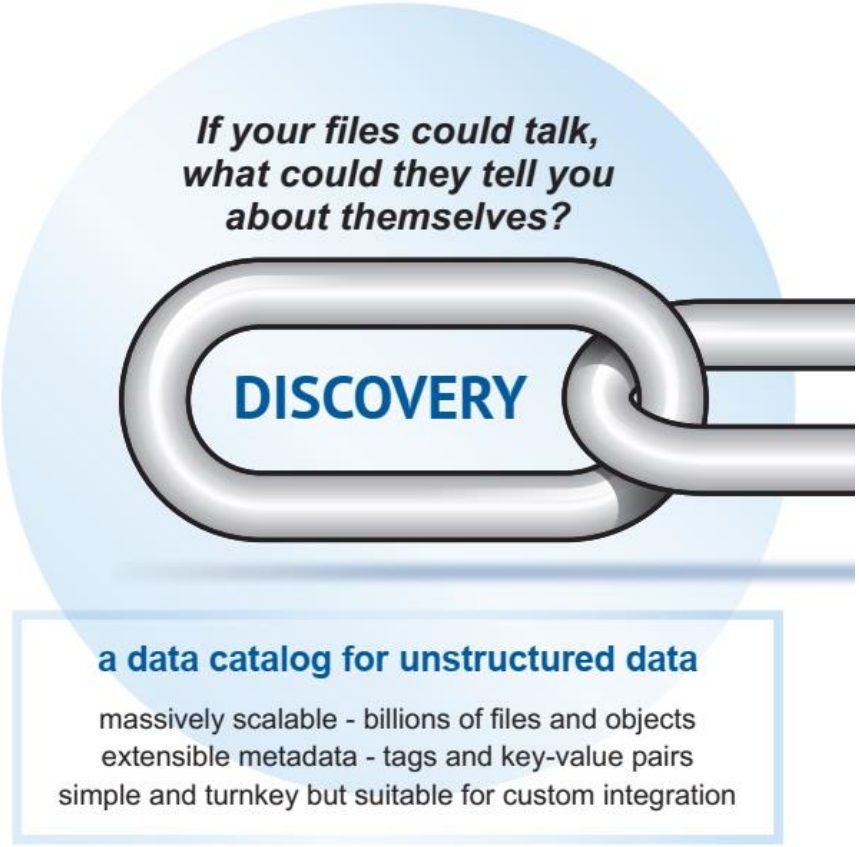


Discovery is the ability to know whatever is knowable about your files, even at very large scale. Use these insights to identify files that you wish to do something to.

Execution is the ability to take action based on your discoveries. Now that you have selected the files, what do you want to do with them?

STARFISH

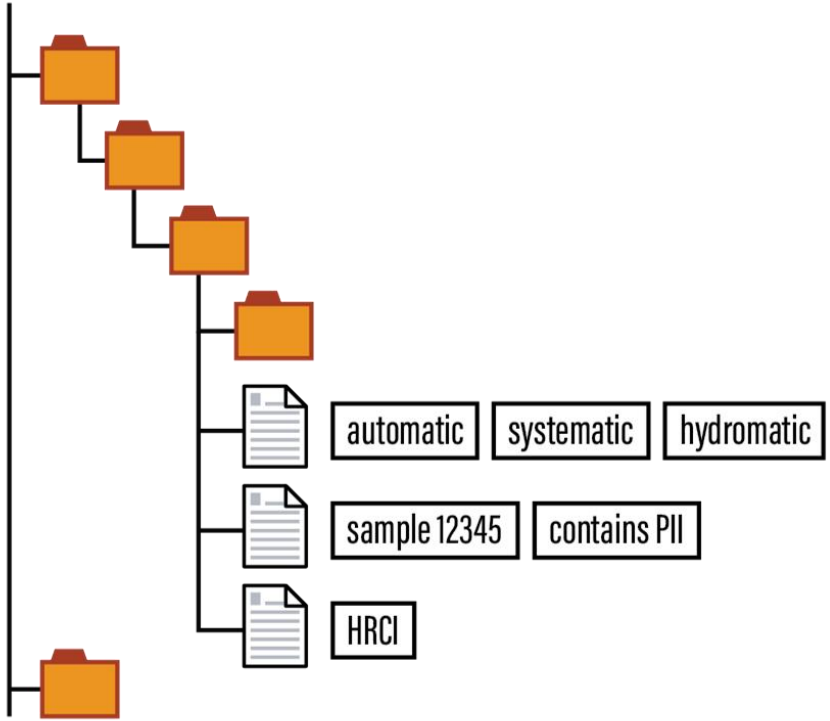
The Discovery Side of Starfish



A PostgreSQL database that enumerates all files and directories.

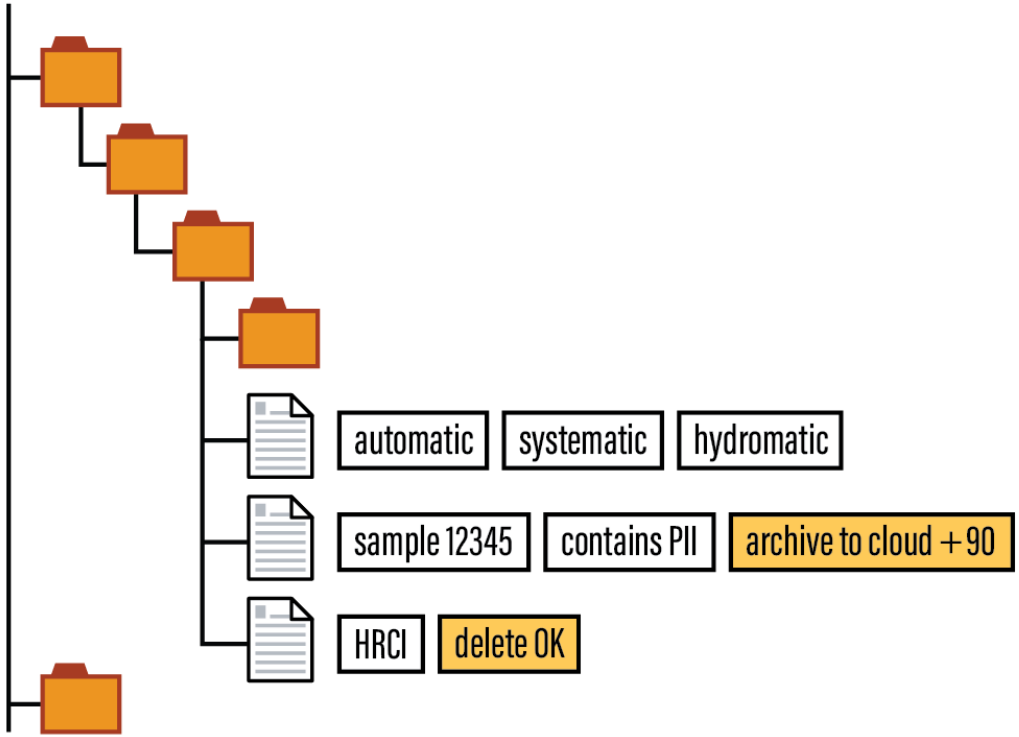
- Retention of the history of the file system metadata over time.
  - Query for specific points in time
  - Enumerate changes in the directory structure and contents between two points in time.
  - Track individual file versions
- Extensible metadata
  - Simple tags
  - Key-value pairs
- Diversity of interfaces
  - CLI, API, HTML5, reporting, dashboards
- Built for very large scale
  - HPC
  - Institutional

Classification Tags: Simple Tagging to Classify Files



- Classification tags are arbitrary strings that add color to files and directories
  - As specific as a unique sample number
  - Regional such as a project code
  - Global such as a general purpose classification
- The same tags can be used across the entire environment
- Tags are typically applied programmatically via API
- Tags are typically selected from a predefined list called a “Tag Set”

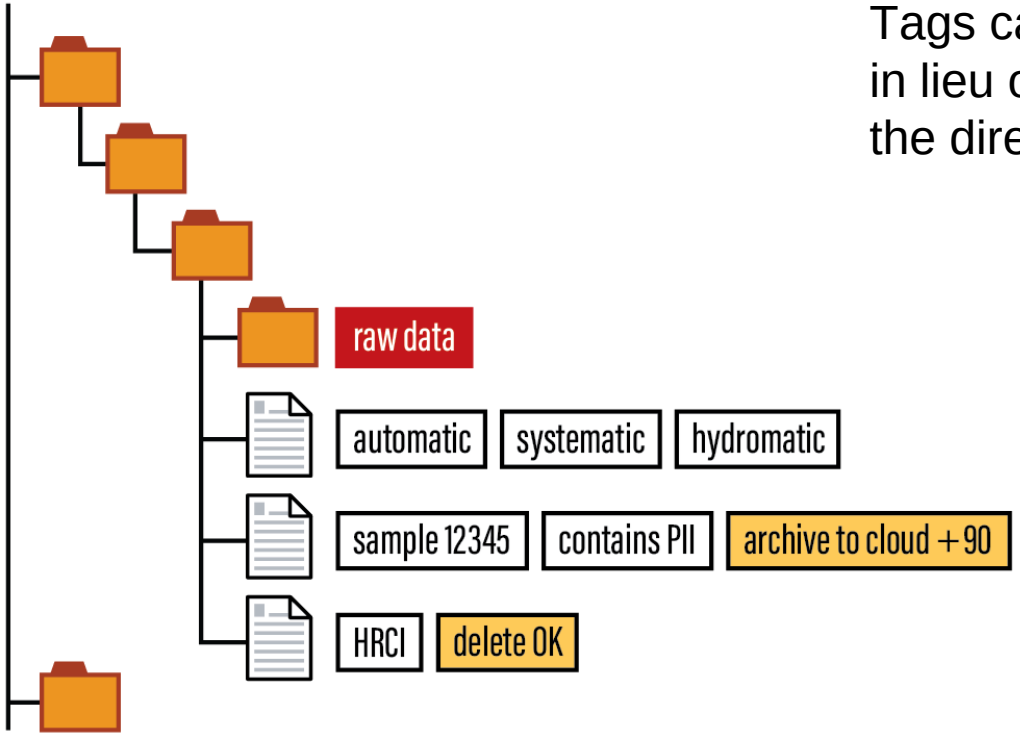
Action Tags Indicate a Desired Action to Take



Action tags are a special kind of tag that denotes that some action is to be taken on the tagged files or objects. Action tags are cleared after the action is taken.

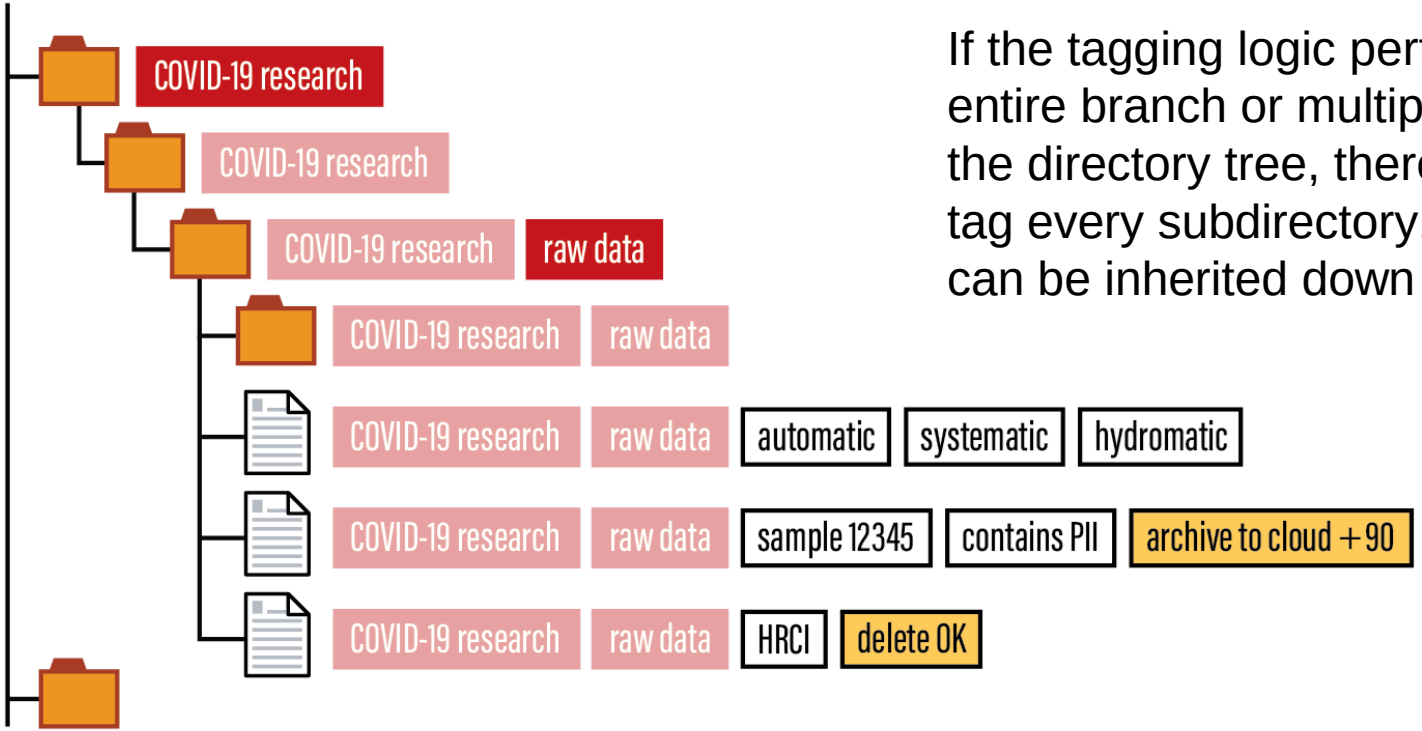
Note that classification tags are also used to identify files upon which to perform actions. The difference is simply whether they are retained or cleared after actions are taken.

Tags Apply to Directories as Well as Individual Files



Tags can be applied at the directory level in lieu of individually tagging every file in the directory.

Directory Tags Are Inherited Down the Branch



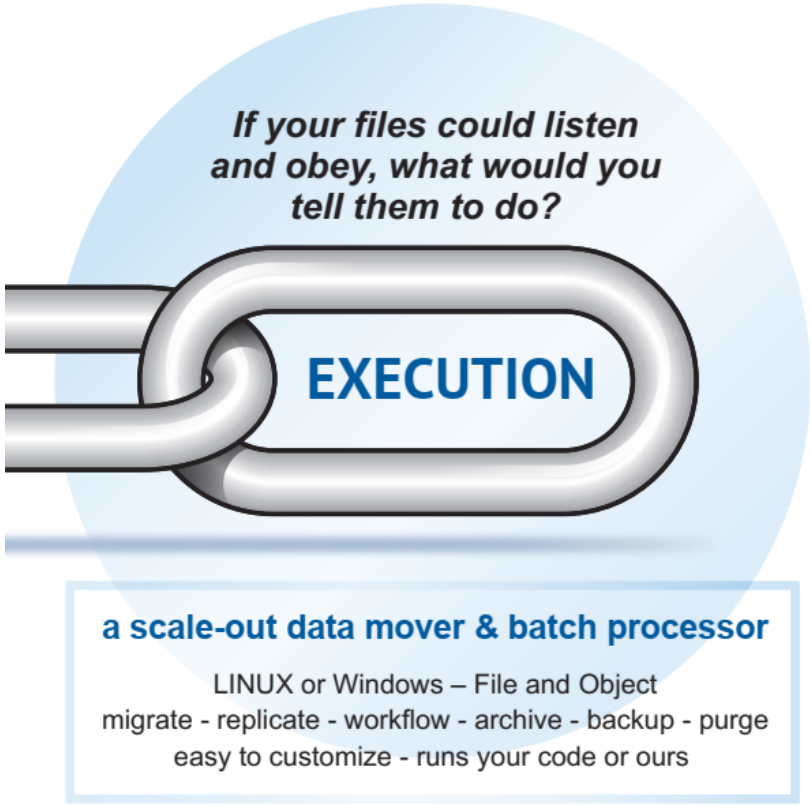
If the tagging logic pertains to an entire branch or multiple branches of the directory tree, there is no need to tag every subdirectory. A single tag can be inherited down the branch.

Best in Class File System Reporting and Analytics

- Starfish is the industry leading solution for reporting and analyzing POSIX-style file systems.
  - Aging, capacity consumption, trending, cost analysis, etc.
- Major differentiators for reporting capabilities are:
  - **Scale** – Support for scale, complexity, and diversity of file systems
    - Products built for the conventional enterprise, simply don't work in scientific computing.
  - **Version history** – few reporting solutions retain essential history, limiting the kinds of reports and insights they can deliver
  - **Metadata** – Starfish's metadata system is unique. Without extensible metadata, file system reports typically lack actionable meaning
  - **SQL openness** – Our database is PostgreSQL, so it is easy to develop new reports or even to expose the database to 3rd party BI tools.
  - **Actionable** – Starfish allows you to take action on on your discoveries.

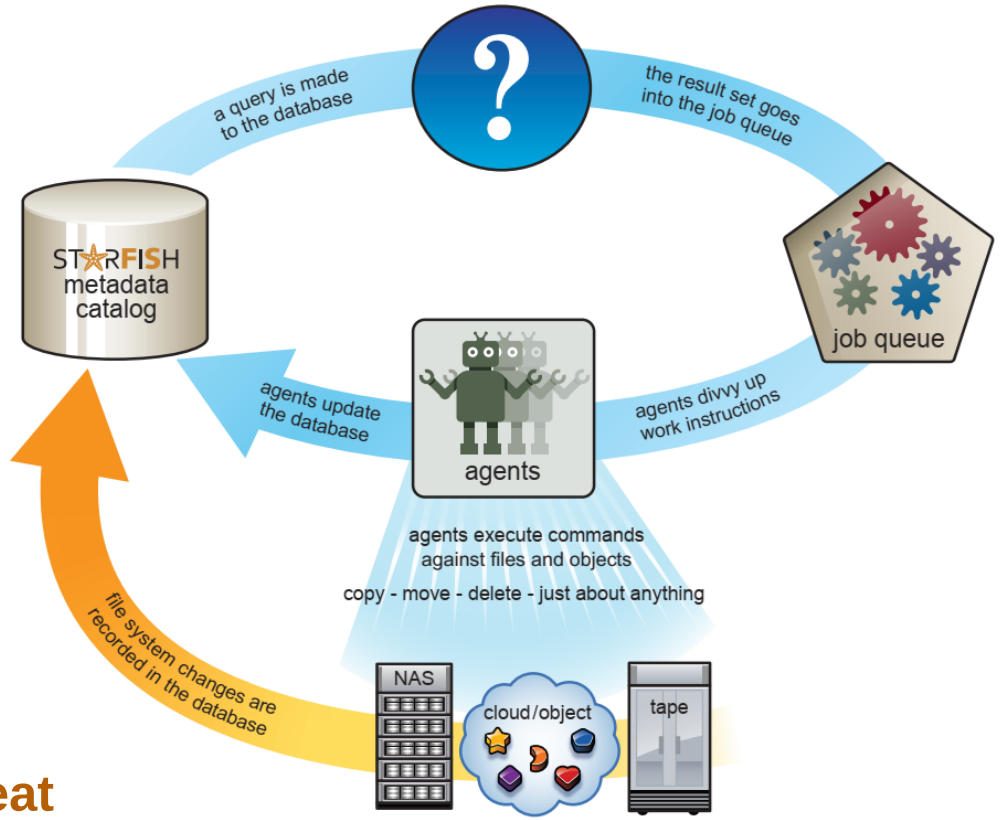
The Execution Side of Starfish: Data Mover / Batch Processor

- The **output** of a query to the data catalog is the **input** for a batch process.
- The batch processor invokes any code against the list of files.
  - Starfish provides code for common functions used across our customer base.
  - You can write your own code in your preferred language.
  - You can execute code from 3rd party providers.
- Batch operations run in parallel across as many servers and threads as is needed for performance objectives.
  - File content or header analysis
  - File copy and move
  - File disposition
- Agents run in either Windows or Linux.



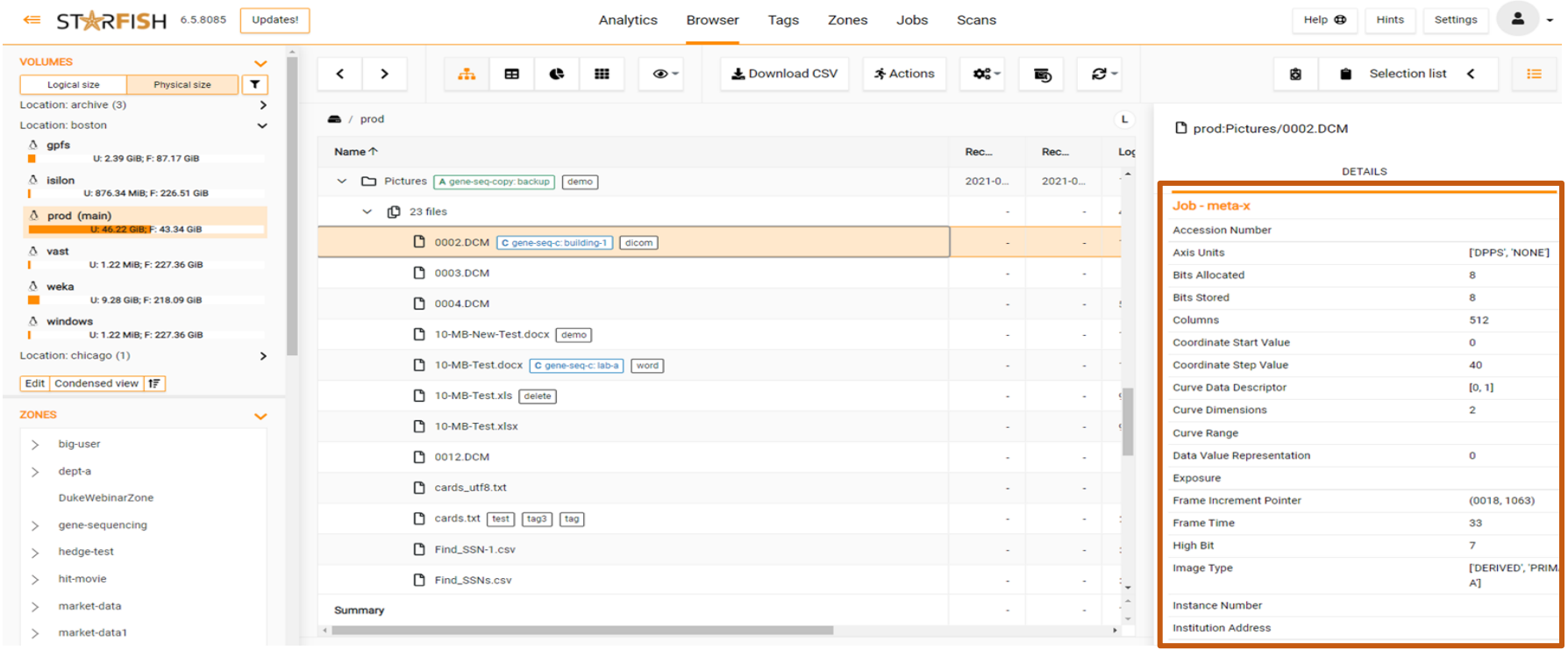
The Feedback Loop Between Catalog and Batch Processor

- The **output** of a query is the **input** of a batch process
- The **output** of the batch process **updates** the database with a **key-value** pair document that describes what was:
  - Done to the file, and/or
  - Discovered about the file
- These **key-value** pairs can be part of the query that defines the next job.

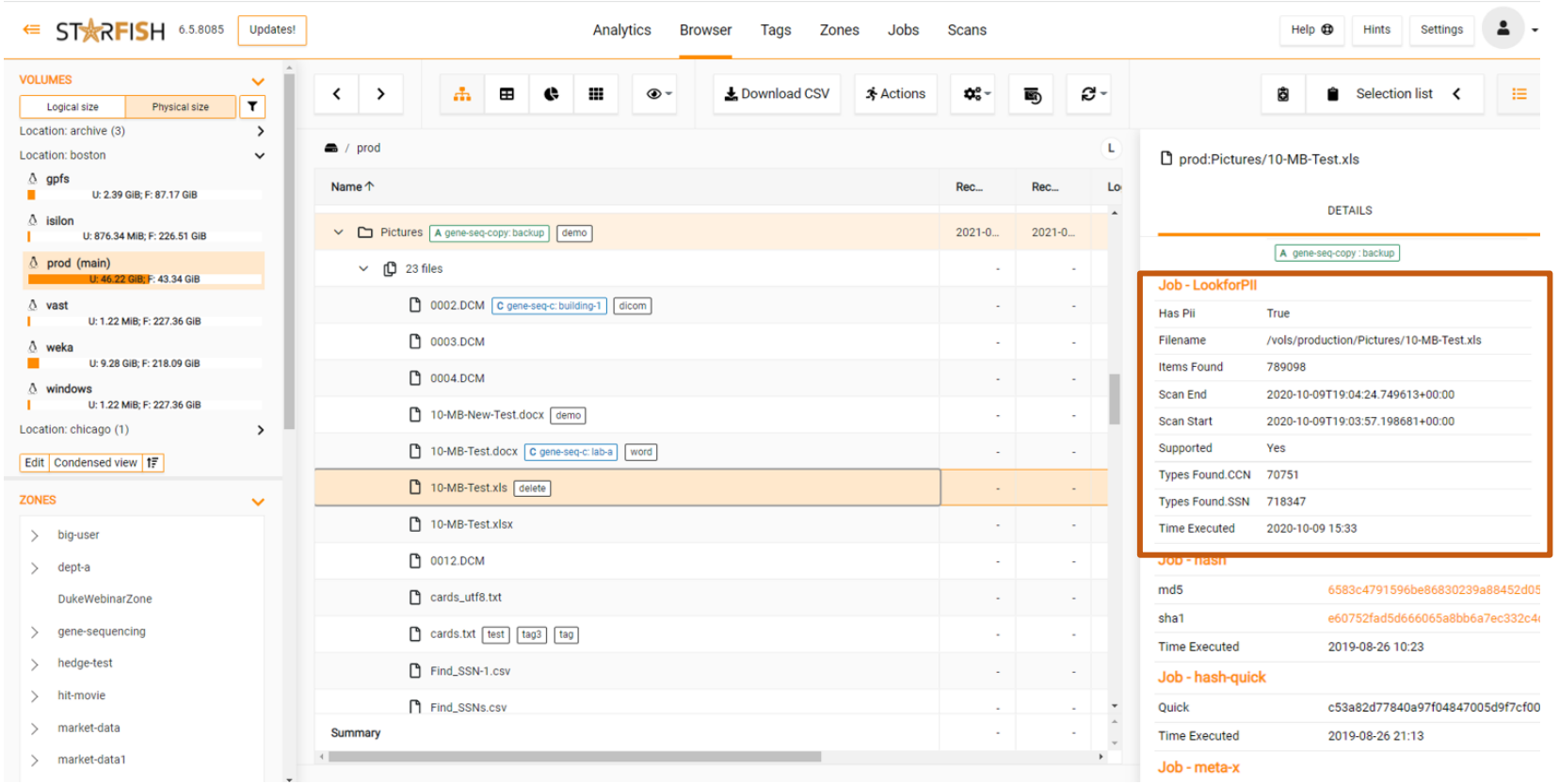


Query - Execute - Update - Repeat

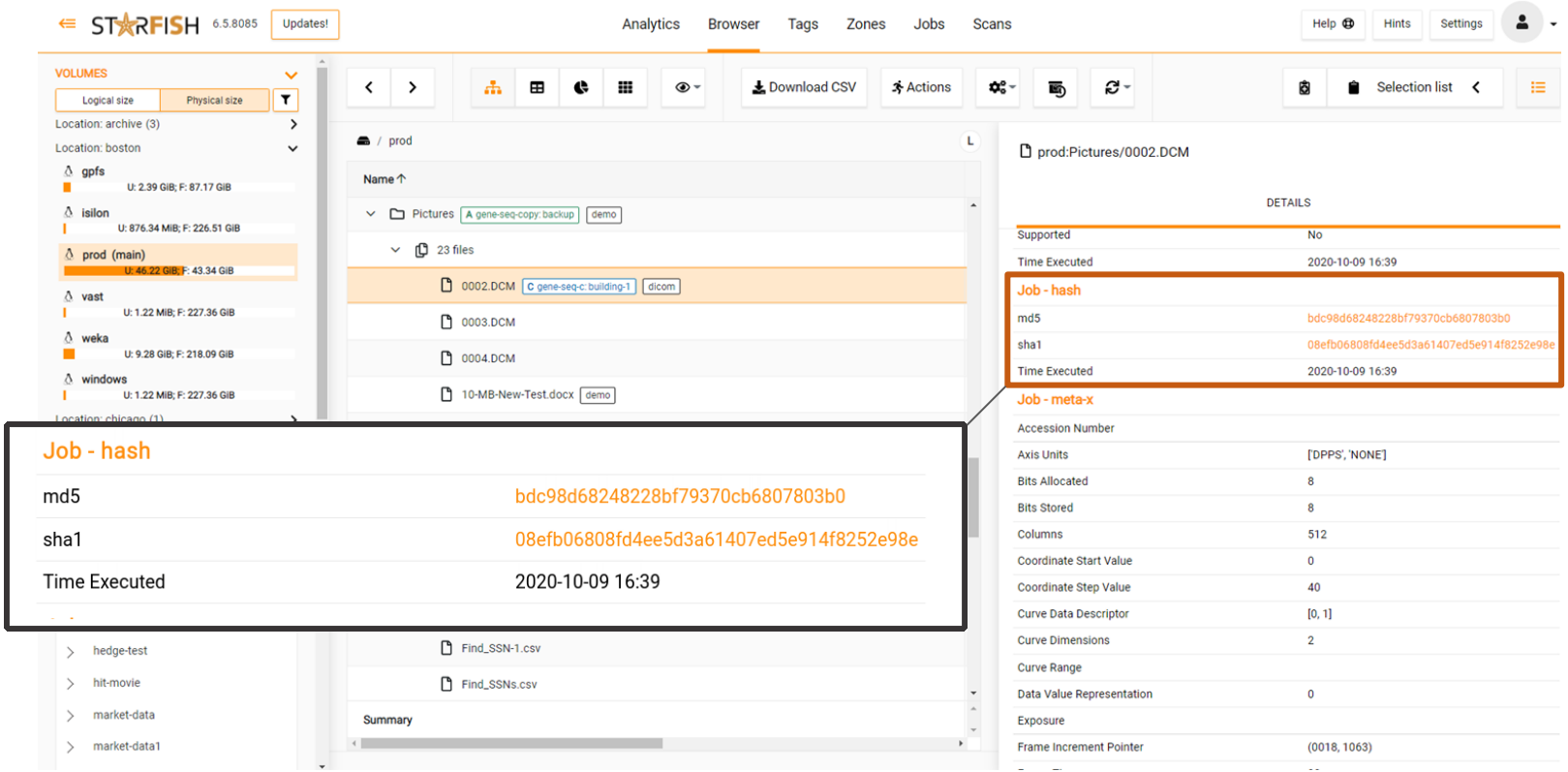
Example of Key-Value Job Results: Metadata Extraction



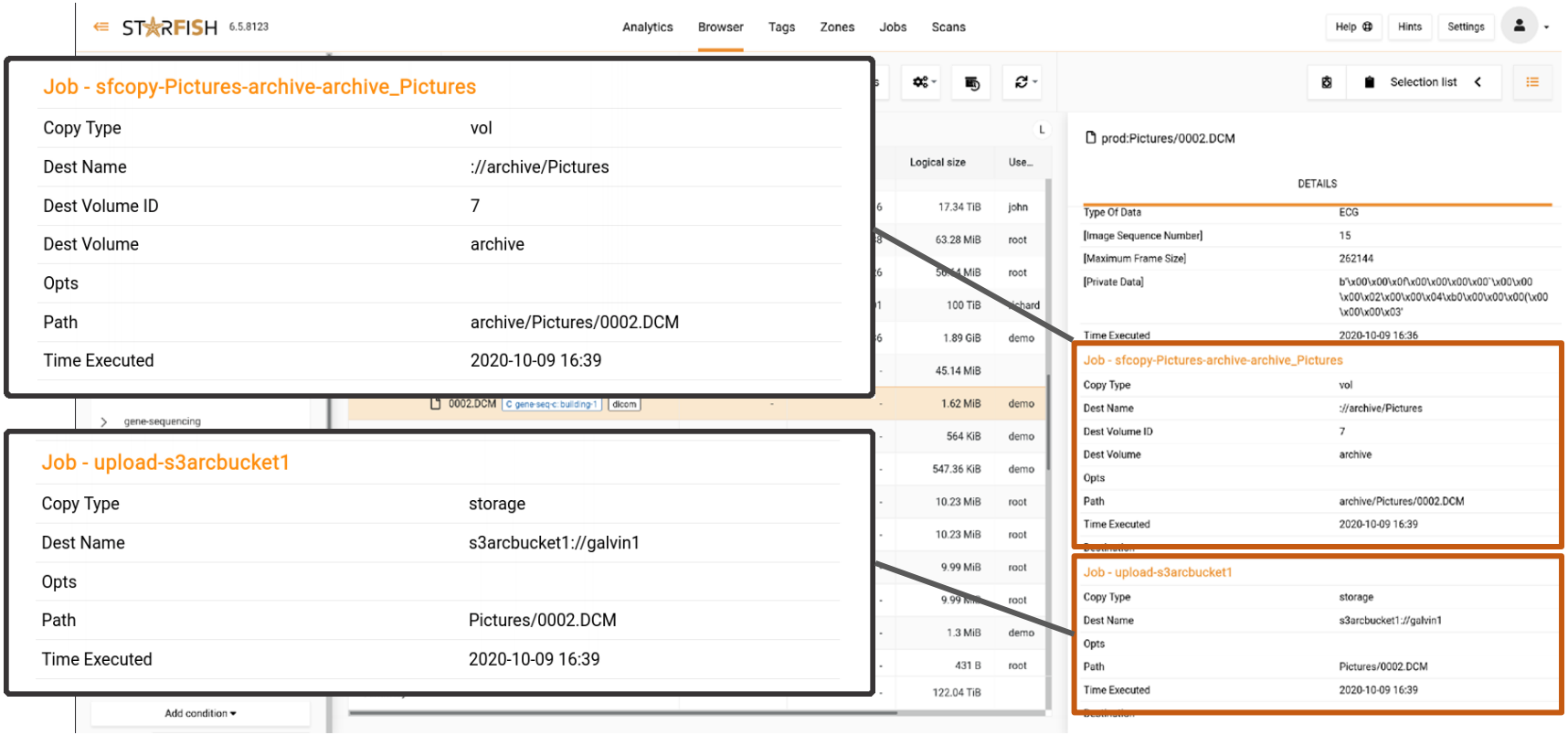
Example of Key-Value Job Results - PII Content Analysis



Example of Key-Value Job Results: Hash Calculation



Example of Key-Value Job Results: Data Movement



Starfish Topology