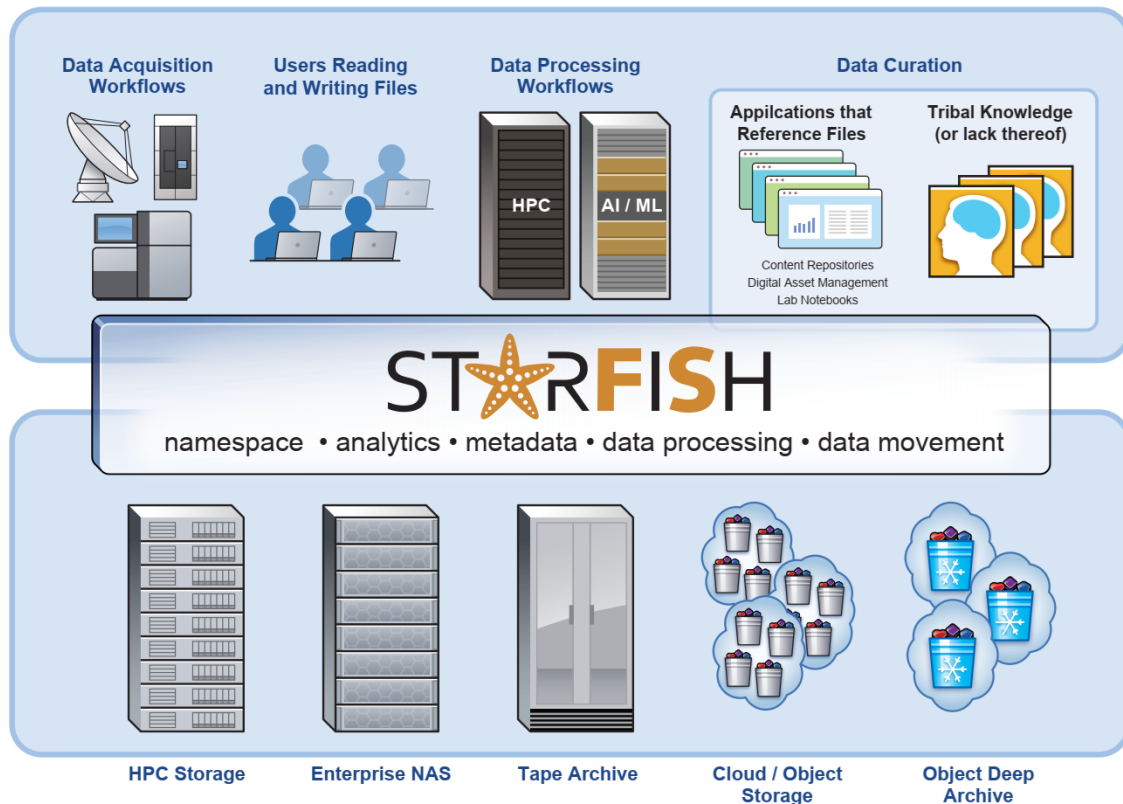


STARFISH

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.

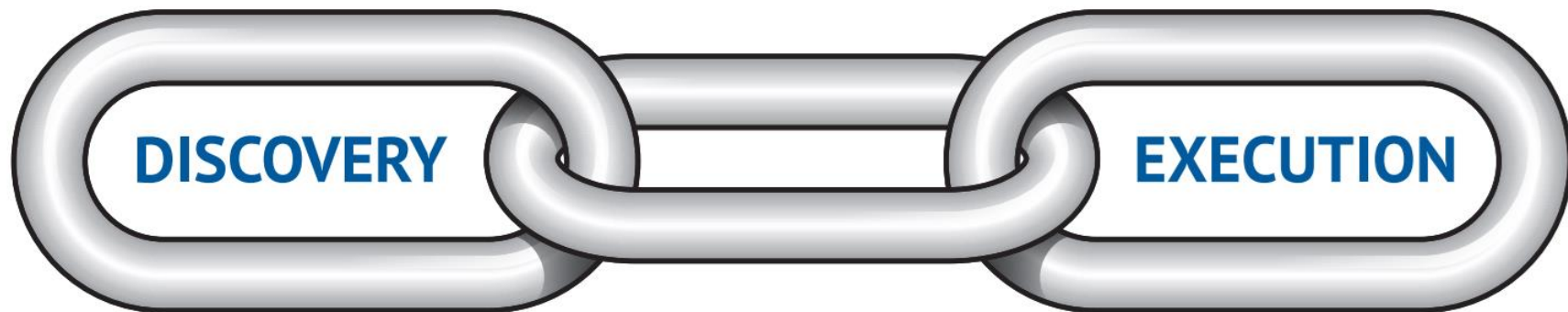
Data curation begins when files are first created.

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?

The Discovery Side of Starfish

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



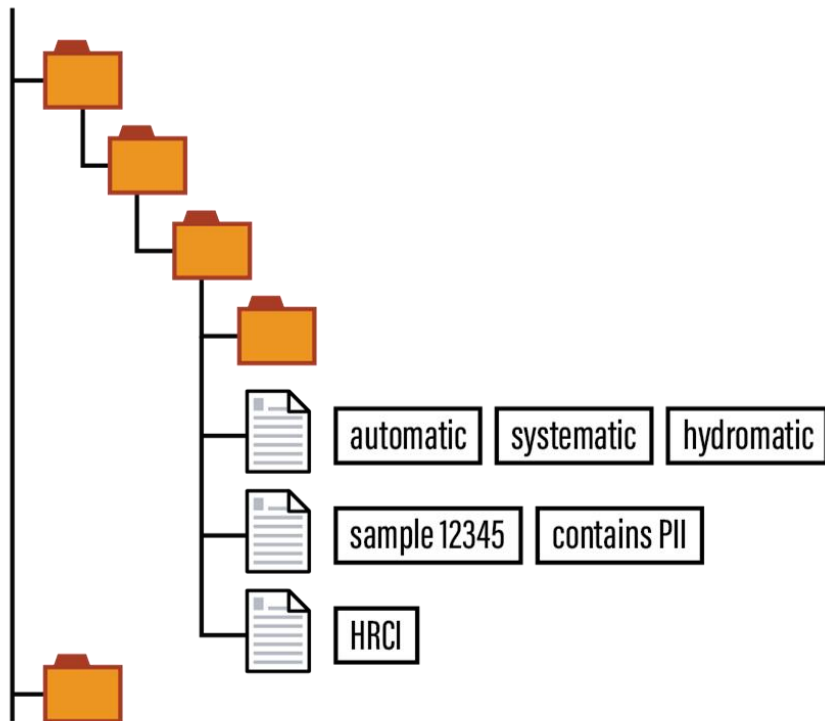
a data catalog for unstructured data

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

A PostgreSQL database that enumerates all files and directories.

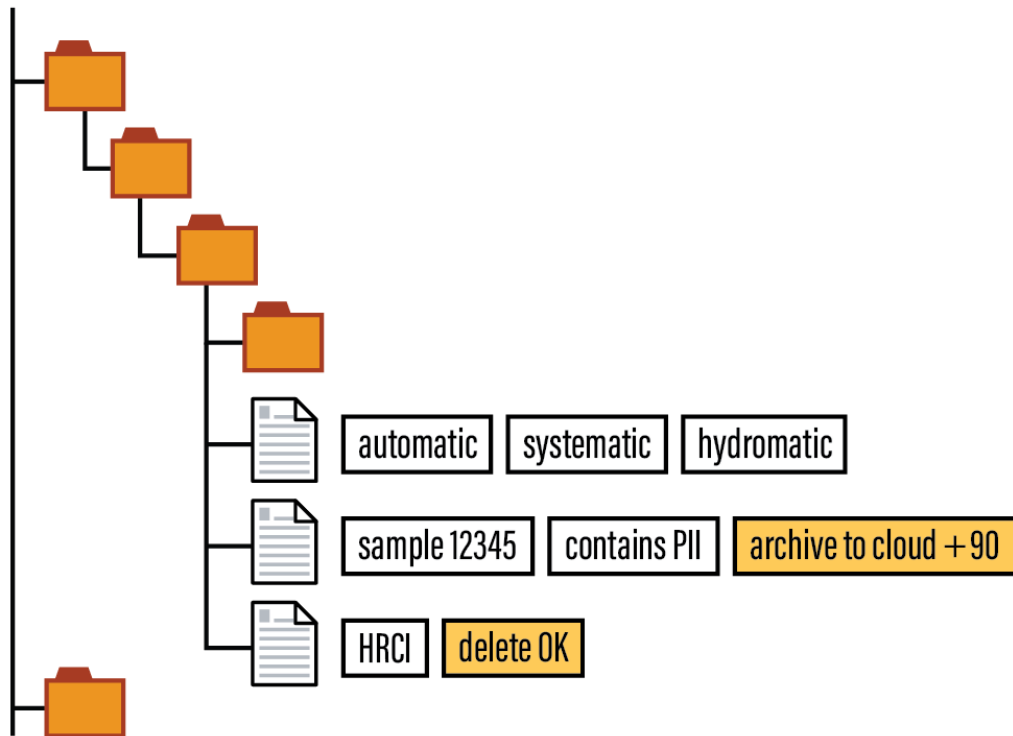
- Retention of the history of the file system metadata over time.
 - Query for specific points in time
 - Enumerate changes in the directory structure and contents between two points in time.
 - Track individual file versions
- Extensible metadata
 - Simple tags
 - Key-value pairs
- Diversity of interfaces
 - 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”

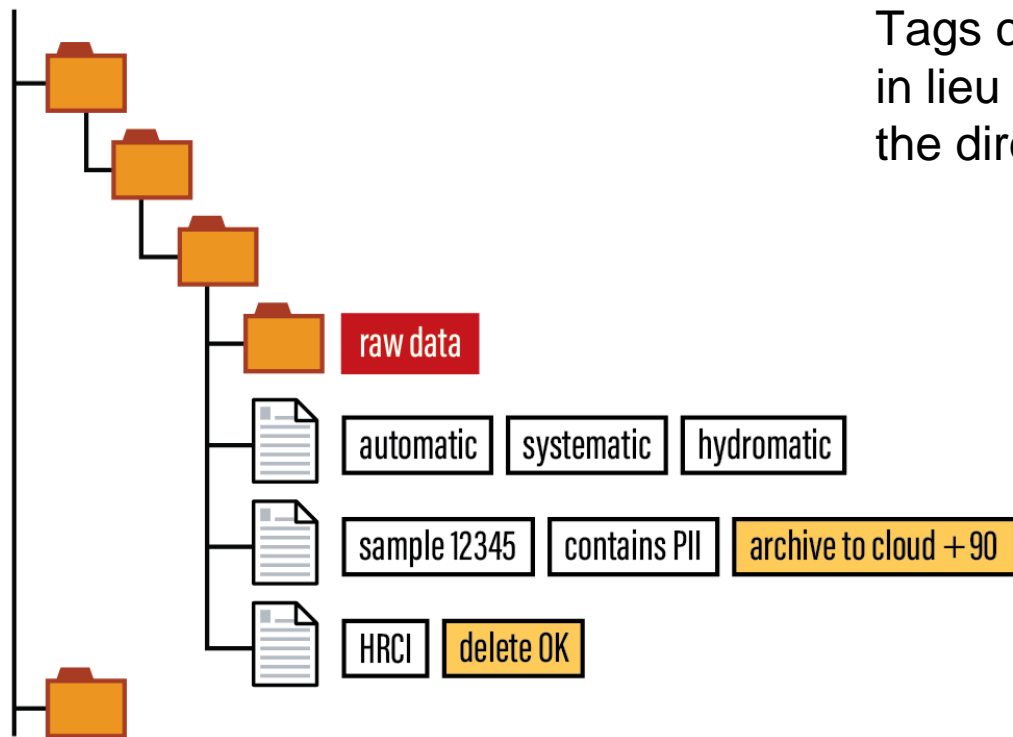
Action Tags Indicate a Desired Action to Take



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.

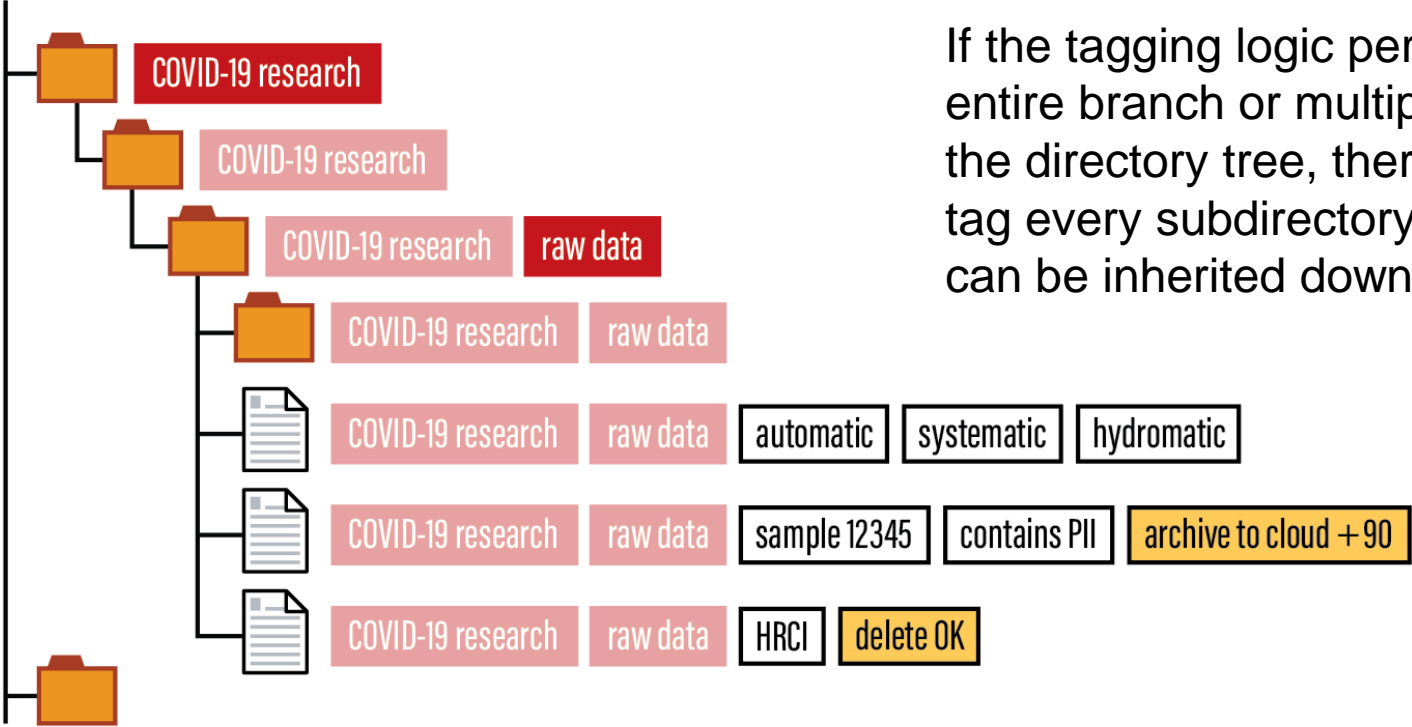
Note that classification tags are also used to identify files upon which to perform actions. The difference is simply whether they are retained or cleared after actions are taken.

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



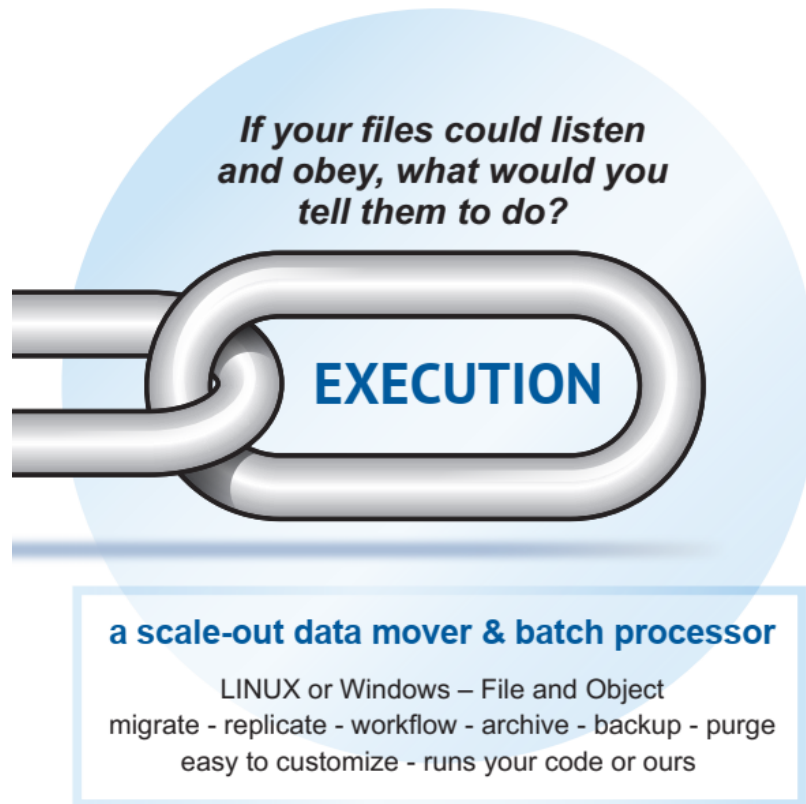
If the tagging logic pertains to an entire branch or multiple branches of the directory tree, there is no need to tag every subdirectory. A single tag can be inherited down the branch.

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.

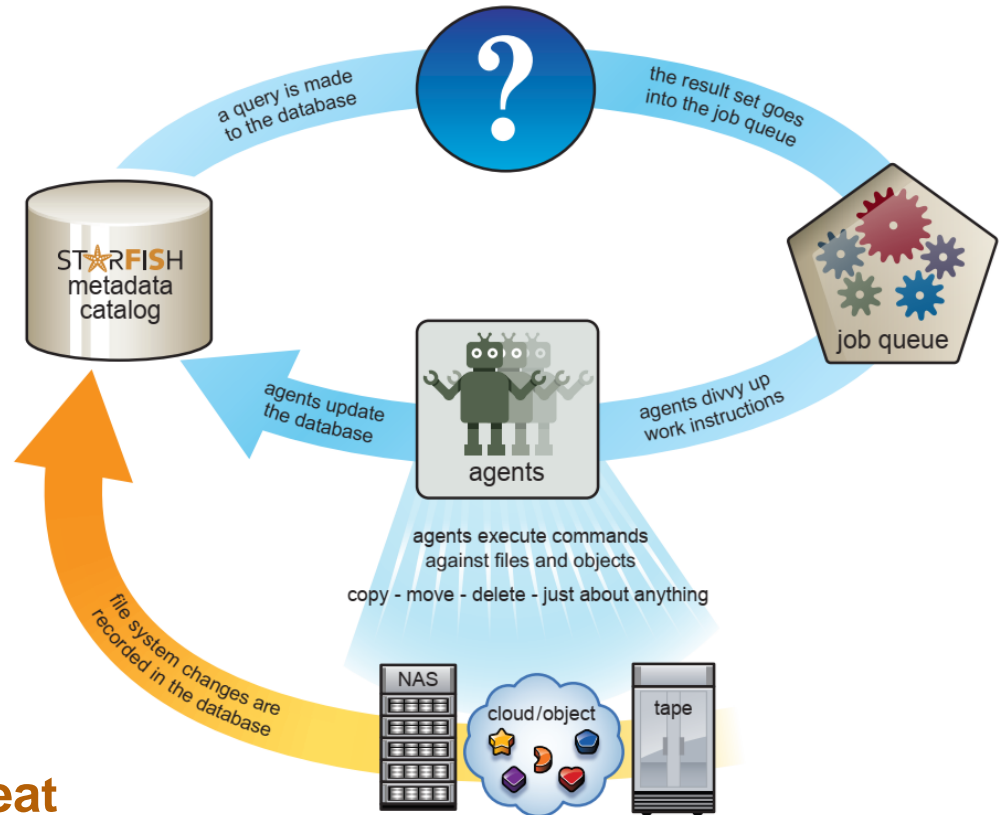
The Execution Side of Starfish: Data Mover / Batch Processor

- 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

The screenshot displays the STARFISH web interface. On the left, there are navigation panels for 'VOLUMES' and 'ZONES'. The 'VOLUMES' panel shows 'prod (main)' selected. The 'ZONES' panel lists various zones like 'big-user', 'dept-a', and 'gene-sequencing'. The main area is a file browser showing a directory structure: 'prod' > 'Pictures' > 'A gene-seq-copy; backup' > 'demo'. A file named '0002.DCM' is selected, with tags 'C gene-seq-c; building-1' and 'dicom'. To the right, a 'DETAILS' panel for 'prod:Pictures/0002.DCM' is shown, containing a 'Job - meta-x' section with various metadata fields.

Name ↑	Rec...	Rec...	Log
▼ Pictures			
▼ A gene-seq-copy; backup	2021-0...	2021-0...	
▼ demo			
23 files			
0002.DCM [C gene-seq-c; building-1] dicom	-	-	
0003.DCM	-	-	
0004.DCM	-	-	
10-MB-New-Test.docx demo	-	-	
10-MB-Test.docx [C gene-seq-c; lab-a] word	-	-	
10-MB-Test.xls delete	-	-	
10-MB-Test.xlsx	-	-	
0012.DCM	-	-	
cards_utf8.txt	-	-	
cards.txt test tag3 tag	-	-	
Find_SSN-1.csv	-	-	
Find_SSNs.csv	-	-	
Summary	-	-	

DETAILS	
Job - meta-x	
Accession Number	
Axis Units	[DPSS, NONE]
Bits Allocated	8
Bits Stored	8
Columns	512
Coordinate Start Value	0
Coordinate Step Value	40
Curve Data Descriptor	[0, 1]
Curve Dimensions	2
Curve Range	
Data Value Representation	0
Exposure	
Frame Increment Pointer	(0018, 1063)
Frame Time	33
High Bit	7
Image Type	[DERIVED, PRIM A]
Instance Number	
Institution Address	

Example of Key-Value Job Results - PII Content Analysis

The screenshot displays the STARFISH interface for a file analysis job. The main view shows a file browser for the path `prod/Pictures/10-MB-Test.xls`. The file list includes various documents and spreadsheets, with `10-MB-Test.xls` highlighted. On the right, a 'DETAILS' panel for the selected file shows the following job information:

Job - LookforPII	
Has PII	True
Filename	/vols/production/Pictures/10-MB-Test.xls
Items Found	789098
Scan End	2020-10-09T19:04:24.749613+00:00
Scan Start	2020-10-09T19:03:57.198681+00:00
Supported	Yes
Types Found.CCN	70751
Types Found.SSN	718347
Time Executed	2020-10-09 15:33

Below the job summary, there are sections for hash and quick analysis:

Job - hash	
md5	6583c4791596be6830239a88452d05
sha1	e60752fad5d666065a8bb6a7ec332c4r
Time Executed	2019-08-26 10:23

Job - hash-quick	
Quick	c53a82d77840a97f04847005d9f7cf00
Time Executed	2019-08-26 21:13

At the bottom, there is a section for meta-analysis:

Job - meta-x

Example of Key-Value Job Results: Hash Calculation

The screenshot shows the STARFISH 6.5.8085 interface. The top navigation bar includes 'Analytics', 'Browser', 'Tags', 'Zones', 'Jobs', and 'Scans'. The left sidebar lists storage volumes: 'prod (main)' with 46.22 GIB U and 43.34 GIB F. The central pane shows a file browser for 'prod/Pictures/0002.DCM'. The right pane displays details for 'prod:Pictures/0002.DCM'. Two callout boxes highlight the 'Job - hash' results:

Job - hash	
md5	bdc98d68248228bf79370cb6807803b0
sha1	08efb06808fd4ee5d3a61407ed5e914f8252e98e
Time Executed	2020-10-09 16:39

Job - meta-x	
Accession Number	
Axis Units	[DPSS', NONE]
Bits Allocated	8
Bits Stored	8
Columns	512
Coordinate Start Value	0
Coordinate Step Value	40
Curve Data Descriptor	[0, 1]
Curve Dimensions	2
Curve Range	
Data Value Representation	0
Exposure	
Frame Increment Pointer	(0018, 1063)

Example of Key-Value Job Results: Data Movement

The screenshot displays the STARFISH 6.5.8123 interface with several job details panels overlaid on a central file browser view.

Job - sfscopy-Pictures-archive-archive_Pictures

Copy Type	vol
Dest Name	://archive/Pictures
Dest Volume ID	7
Dest Volume	archive
Opts	
Path	archive/Pictures/0002.DCM
Time Executed	2020-10-09 16:39

Job - upload-s3arcbucket1

Copy Type	storage
Dest Name	s3arcbucket1://galvin1
Opts	
Path	Pictures/0002.DCM
Time Executed	2020-10-09 16:39

Job - sfscopy-Pictures-archive-archive_Pictures (Details)

Type Of Data	ECG
[Image Sequence Number]	15
[Maximum Frame Size]	262144
[Private Data]	b'x00x00x0f0x00x00x00x00'x00x00 x00x02x00x00x004'xb0'x00x00'x00 x00x00x03'
Time Executed	2020-10-09 16:36

Job - upload-s3arcbucket1 (Details)

Copy Type	storage
Dest Name	s3arcbucket1://galvin1
Opts	
Path	Pictures/0002.DCM
Time Executed	2020-10-09 16:39

The background interface shows a file browser with columns for Logical size and Use... and a table listing files with their sizes and users. A file named '0002.DCM' is highlighted.

Starfish Topology

