



Tiering and Life Cycle Management with AI/ML Workloads

MSST - 2019

Jacob Farmer

- CTO, Cambridge Computer
- Chief Evangelist, Starfish Storage

Thesis Statement:

Machine learning creates demand for new classes of storage and thus provides impetus for the adoption of better practices for data life cycle management.

When are Tiering and Lifecycle Management Most Interesting?



- When there is a wide spread in costs between the fastest tiers and the cheapest tiers
- When the data has a life cycle or usage pattern that allows for meaningful savings when files can be moved between tiers.
 - When the cost and complexity of data life cycle management does not outweigh the savings of putting all data on a single tier



• Suitable performance is a **MUST HAVE**

- Compute resources are expensive and sit idle when the storage system fails to bring the data

• There are tons of new solutions

- Super fast, massively scalable flash storage systems
- Specialized software interfaces for bypassing the kernel to load GPUs
- In memory compute using capacity optimized RAM disks

• How much top tier do you need?

- How do you measure?

• How do you avoid wasting top tier capacity?

- Stale data should not sit in expensive storage

- **The main purpose of the secondary tiers is to swap files with the primary tier**
 - They need to be optimized for suitable data transfer performance
- **This is different from typical HPC life cycle management**
 - In conventional HPC many workloads are happy on a middle level tier
 - Many conventional workloads can take the latency hit reading the file from a lower tier while promoting the bit to a higher tier.

- Most large datasets in science fall into the **WORSE** or **WORN** category:
 - WORN – Write Once Read Never
 - WORSE – Write Once Read Seldom if Every
- Machine learning sets are much more likely to re-used and when they are reused, very large data sets need to be retrieved from archive.

In short:

Machine learning workloads require more aggressive staging and de-staging between tiers than traditional scientific computing workloads.

There are Many Solutions for Federating Tiers of Storage

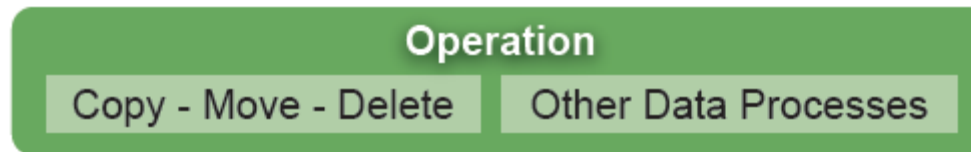
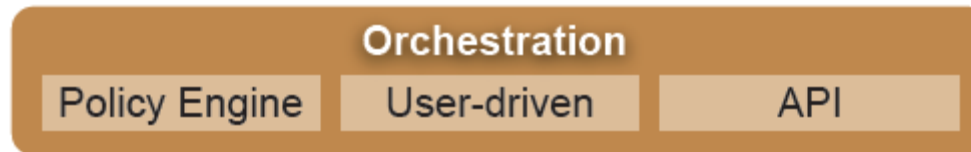


- **File systems with multiple tiers**
 - Newer file systems leverage SSDs and cloud tiers natively
 - Many file systems can subsume external storage devices and incorporate them into their name space.
- **PNFS is back and can provide a global namespace across multiple devices**
- **You can simply have multiple storage devices (even local staging disk), each with their own namespace and move files yourself.**
 - Logical namespace in middleware
 - Logical namespace in application software



- Where is my file?
- Where do I want it to be instead?
- Is it there yet?

The OOO Model For Data Life Cycle Management



- **A unified POSIX-like namespace is perhaps less important because Machine Learning is driven by the machine.**
 - The machine does not need a friendly pathname in a hierarchical structure
- **The workflow will likely be driven by metadata stored in an application**

- **You can't rely on file touches to trigger migration.**
 - You have to be able to stage and de-stage in advance
- **What are you going to do? Chances are you will do the following (because this is what everyone pretty much does)**
 - Make a database of your files
 - Add metadata to your database to make it easy to specify which files you want
 - Query the database to generate a migration script
 - Run the script on a scheduled basis or integrate with job scheduler



CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

Starfish (*FS)

**A Software Company Spun Out of
Cambridge Computer**

- **Starfish makes and maintains a database of your file system**
- **Starfish allows you to associate metadata with files and directories**
 - Gather metadata from file system, individual files, from the workflow, or wherever.
 - Copy select metadata from application software
- **Starfish uses the query result to feed batch processor that executes code against the files**
 - Batch processing runs in parallel across multiple agents
 - Agents are ordinary LINUX machines (Windows agent later this year).
 - You can borrow nodes from the compute farm
- **Of course, all of this is API driven, easy to use, and feature rich**

What Makes Our Database Implementation Special



- **It is open. We use PostGres.**
- **We handle extreme scale**
 - Billions of files
 - Thousands of change events per second
- **File and Directory Metadata**
 - Simple tags on files and directories (inheritable or not)
 - Key-value pairs for individual files
- **We keep the version history of the directory tree and of individual files**
- **We aggregate values for lightning fast insights**
- **We take action on the query results**
 - COPY, MOVE, DELETE, GET, PUT, etc.



Starfish is Made Up of Three Main Components



• Core database

- We synchronize the metadata in your POSIX file systems with a database.
- We allow additional metadata to be added to files and directories

• Jobs engine

- A batch processor that takes the results of the query and does “stuff”.
 - Copy, move, delete
 - Calculate hashes
 - Extract metadata
 - Your code or ours
- Work is divvied up among any number of agents

• User Interface

- HTML-5 file system browser
- Discovery and system monitoring
- (Beta) User portal that allows users to participate in storage management policies



Search / Query Builder



STARFISH Developer Preview Dashboard Browser Jobs Scans Tags Administration

740 search results (6 dirs and 734 files) matching "edic*"

File	Size	Owner	Count	Cost	Modified	Accessed
Biomedicine_(Taipei)	6.63 MiB	john	151	0.65	2017-05-23 10:34	2017-09-28 19:27
Biomedicine_(Taipei)_2014_Aug_13_4_16.txt	19.3 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_13_4_18.txt	16.81 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_13_4_19.txt	30.07 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_13_4_20.txt	27.6 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_14_11.txt	20.13 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_25_4_21.txt	31.68 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_27_4_17.txt	27.99 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_27_4_3.txt	23.22 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_27_4_5.txt	23.13 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_2_4_10.txt	23.08 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_2_4_14.txt	17.04 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_4_4_12.txt	39.44 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_5_4_2.txt	54.57 KiB	john	1	0.01	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_6_4_4.txt	19.07 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_6_4_6.txt	24.33 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_6_4_7.txt	14.27 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_6_4_8.txt	34.27 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Aug_6_4_9.txt	30.91 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Feb_12_4_1.txt	34.66 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Feb_3_4_13.txt	33.55 KiB	john	1	0.00	2016-07-12 18:45	2017-09-28 19:18
Biomedicine_(Taipei)_2014_May_8_4_15.txt	13.28 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Nov_13_4_25.txt	63.68 KiB	john	1	0.01	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Nov_16_4_24.txt	47.37 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Nov_20_4_23.txt	44.71 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Nov_22_4_26.txt	18.6 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Nov_26_4_27.txt	8.03 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2014_Nov_26_4_22.txt	36.62 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Biomedicine_(Taipei)_2015_Aun 11 5(3) 13.txt	50.03 KiB	john	1	0.00	2016-07-12 18:41	2017-09-28 19:18
Summary	13.09 GiB		740	1309.19		

Storage

- Has copy Has been archived or backed up
- Latest copied version Only the latest version, beyond primary storage
- Number of hardlinks to file Number
- On disk On primary storage (or only in secondary)

Metadata

- Job id Starfish Job ID
- Jobs Starfish job results
- Tag (explicit) Starfish tag applied directly to item
- Tag (inherited) Starfish tag applied directly or above in the directory tree

CURRENT VIEW FILTERS

- ACCESS TIME >10 years
- MODIFICATION TIME >10 years
- DIRECTORY SIZE RECURSIVE max: 30.9 GiB

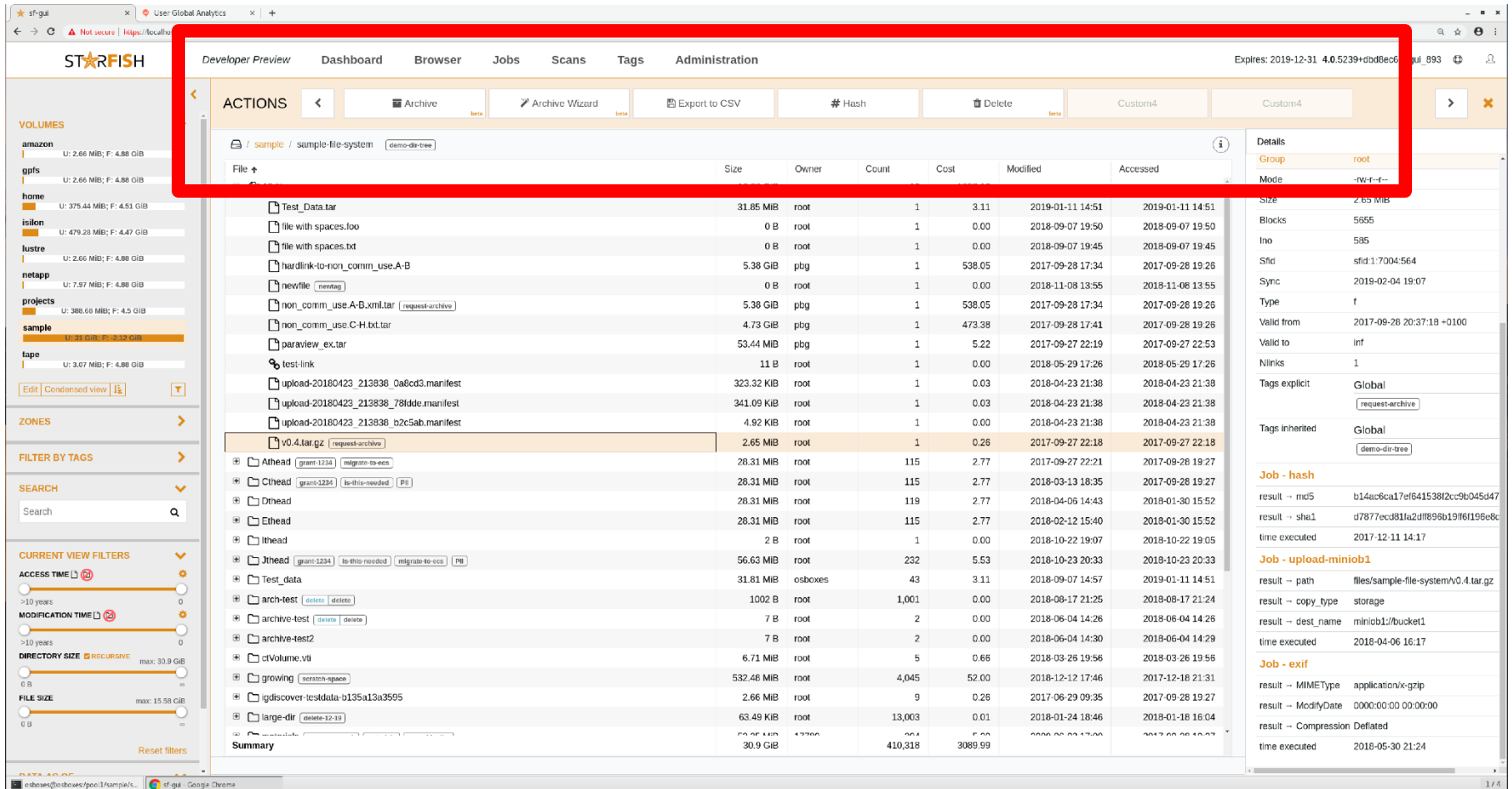


Metadata / Details



The screenshot displays the STARFISH web interface. On the left, there are navigation panels for 'VOLUMES' (listing various storage volumes like amazon, gpps, home, etc.), 'ZONES', 'FILTER BY TAGS', 'SEARCH', and 'CURRENT VIEW FILTERS'. The main area shows a file browser view for 'sample / sample-file-system'. A table lists 13 files with columns for File, Size, Owner, Count, Cost, Modified, and Accessed. The file 'v0.4.tar.gz' is selected and highlighted in orange. On the right, a 'Details' panel is open, showing metadata for the selected file, including Group (root), Mode (-rw-r--r--), Size (2.65 MiB), Blocks (5655), Ino (585), Sfid (sfid:17004564), Sync (2019-02-04 19:07), Type (f), Valid from (2017-09-28 20:37:18 +0100), Valid to (inf), Nlinks (1), Tags explicit (Global), Tags inherited (Global), Job - hash, Job - upload-miniob1, and Job - exif.

File	Size	Owner	Count	Cost	Modified	Accessed
13 files	15.58 GiB		13	1558.13		
Test_Data.tar	31.85 MiB	root	1	3.11	2019-01-11 14:51	2019-01-14:51
file with spaces.foo	0 B	root	1	0.00	2018-09-07 19:50	2018-09-19:50
file with spaces.txt	0 B	root	1	0.00	2018-09-07 19:45	2018-09-18:45
hardlink-to-non_comm_use.A-B	5.38 GiB	pbj	1	538.05	2017-09-28 17:34	2017-09-19:26
newfile (newtag)	0 B	root	1	0.00	2018-11-08 13:55	2018-11-13:55
non_comm_use.A-B.xml.tar (request-archive)	5.38 GiB	pbj	1	538.05	2017-09-28 17:34	2017-09-19:26
non_comm_use.C-H.txt.tar	4.73 GiB	pbj	1	473.38	2017-09-28 17:41	2017-09-19:26
paraview_ex.tar	53.44 MiB	pbj	1	5.22	2017-09-27 22:19	2017-09-22:53
test-link	11 B	root	1	0.00	2018-05-29 17:26	2018-05-17:26
upload-20180423_213838_0u8cd3.manifest	323.32 KiB	root	1	0.03	2018-04-23 21:38	2018-04-21:38
upload-20180423_213838_78kde.manifest	341.09 KiB	root	1	0.03	2018-04-23 21:38	2018-04-21:38
upload-20180423_213838_b2c5ab.manifest	4.92 KiB	root	1	0.00	2018-04-23 21:38	2018-04-21:38
v0.4.tar.gz (request-archive)	2.65 MiB	root	1	0.26	2017-09-27 22:18	2017-09-22:18
Atthead (grant:1234) (migrate-to-ecs)	28.31 MiB	root	115	2.77	2017-09-27 22:21	2017-09-19:27
Cthead (grant:1234) (is-this-needed) (PI)	28.31 MiB	root	115	2.77	2018-03-13 18:35	2017-09-19:27
Dthead	28.31 MiB	root	119	2.77	2018-04-06 14:43	2018-01-15:52
Ethead	28.31 MiB	root	115	2.77	2018-02-12 15:40	2018-01-15:52
Ithead	2 B	root	1	0.00	2018-10-22 19:07	2018-10-19:05
Jthead (grant:1234) (is-this-needed) (migrate-to-ecs) (PI)	56.63 MiB	root	232	5.53	2018-10-23 20:33	2018-10-20:33
Test_data	31.81 MiB	osboxes	43	3.11	2018-09-07 14:57	2019-01-14:51
arch-test (delete) (delete)	1002 B	root	1,001	0.00	2018-08-17 21:25	2018-08-21:24
archive-test (delete) (delete)	7 B	root	2	0.00	2018-06-04 14:26	2018-06-14:26
archive-test2	7 B	root	2	0.00	2018-06-04 14:30	2018-06-14:29
ctVolume.vti	6.71 MiB	root	5	0.66	2018-03-26 19:56	2018-03-19:56
growing (scratchspace)	532.48 MiB	root	4,045	52.00	2018-12-12 17:46	2017-12-21:31
igdiscover-testdata-b135a13a3595	2.66 MiB	root	9	0.26	2017-06-29 09:35	2017-09-19:27
large-dir (delete-12-18)	63.49 KiB	root	13,003	0.01	2018-01-24 18:46	2018-01-16:04
Summary	30.9 GiB		410,318	3089.99		



The screenshot shows the STARFISH web interface. The top navigation bar includes "Developer Preview", "Dashboard", "Browser", "Jobs", "Scans", "Tags", and "Administration". The "ACTIONS" menu is highlighted with a red box, showing options like "Archive", "Archive Wizard", "Export to CSV", "# Hash", "Delete", and "Custom4". Below the menu is a table of files and directories. The table has columns for "File", "Size", "Owner", "Count", "Cost", "Modified", and "Accessed". The "v0.4.tar.gz" file is highlighted in orange. On the right side, there is a "Details" panel showing file metadata such as "Group: root", "Mode: -rwxr-xr-x", "Size: 2.65 MiB", "Blocks: 5655", "Ino: 585", "Sfd: sfd:1:7004:564", "Sync: 2019-02-04 19:07", "Type: f", "Valid from: 2017-09-26 20:37:18 +0100", "Valid to: inf", "Nlinks: 1", "Tags explicit: Global", "Tags inherited: Global", "Job - hash", "Job - upload-miniob1", and "Job - exif".

File	Size	Owner	Count	Cost	Modified	Accessed
Test_Data.tar	31.85 MiB	root	1	3.11	2019-01-11 14:51	2019-01-11 14:51
file with spaces.foo	0 B	root	1	0.00	2018-09-07 19:50	2018-09-07 19:50
file with spaces.txt	0 B	root	1	0.00	2018-09-07 19:45	2018-09-07 19:45
hardlink-to-non_comm_use.A-B	5.38 GiB	pbjg	1	538.05	2017-09-28 17:34	2017-09-28 19:26
newfile (newing)	0 B	root	1	0.00	2018-11-08 13:55	2018-11-08 13:55
non_comm_use.A-B.xml.tar (request-archive)	5.38 GiB	pbjg	1	538.05	2017-09-28 17:34	2017-09-28 19:26
non_comm_use.C-H.txt.tar	4.73 GiB	pbjg	1	473.38	2017-09-28 17:41	2017-09-28 19:26
paraview_ext.tar	53.44 MiB	pbjg	1	5.22	2017-09-27 22:19	2017-09-27 22:53
test-link	11 B	root	1	0.00	2018-05-29 17:26	2018-05-29 17:26
upload-20180423_213838_0a8cd3.manifest	323.32 KiB	root	1	0.03	2018-04-23 21:38	2018-04-23 21:38
upload-20180423_213838_781dde.manifest	341.09 KiB	root	1	0.03	2018-04-23 21:38	2018-04-23 21:38
upload-20180423_213838_b2c5ab.manifest	4.92 KiB	root	1	0.00	2018-04-23 21:38	2018-04-23 21:38
v0.4.tar.gz (request-archive)	2.65 MiB	root	1	0.26	2017-09-27 22:18	2017-09-27 22:18
Athead (grant-1234) (migrate-to-ecs)	28.31 MiB	root	115	2.77	2017-09-27 22:21	2017-09-26 19:27
Cthead (grant-1234) (is-this-needed) (PI)	28.31 MiB	root	115	2.77	2018-03-13 18:35	2017-09-26 19:27
Dthead	28.31 MiB	root	119	2.77	2018-04-06 14:43	2018-01-30 15:52
Ethead	28.31 MiB	root	115	2.77	2018-02-12 15:40	2018-01-30 15:52
Ithead	2 B	root	1	0.00	2018-10-22 19:07	2018-10-22 19:05
Jthead (grant-1234) (is-this-needed) (migrate-to-ecs) (PI)	56.63 MiB	root	232	5.53	2018-10-23 20:33	2018-10-23 20:33
Test_data	31.81 MiB	osboxes	43	3.11	2018-09-07 14:57	2019-01-11 14:51
arch-test (delete) (delete)	1002 B	root	1,001	0.00	2018-08-17 21:25	2018-08-17 21:24
archive-test (delete) (delete)	7 B	root	2	0.00	2018-06-04 14:26	2018-06-04 14:26
archive-test2	7 B	root	2	0.00	2018-06-04 14:30	2018-06-04 14:29
ctVolume.vti	6.71 MiB	root	5	0.66	2018-03-26 19:56	2018-03-26 19:56
growing (scratch-space)	532.48 MiB	root	4,045	52.00	2018-12-12 17:46	2017-12-18 21:31
igdiscover_tcsdata_b135a13a3595	2.66 MiB	root	9	0.26	2017-06-29 09:35	2017-09-28 19:27
large-dir (delete-12-13)	63.49 KiB	root	13,003	0.01	2018-01-24 18:46	2018-01-18 16:04
Summary	30.9 GiB		410,318	3089.99		



A "Virtual" Global File System

