



Access Controlled AI Context with GUFi

Jason Lee, Gary Grider

June 1, 2026

LA-UR-26-24285

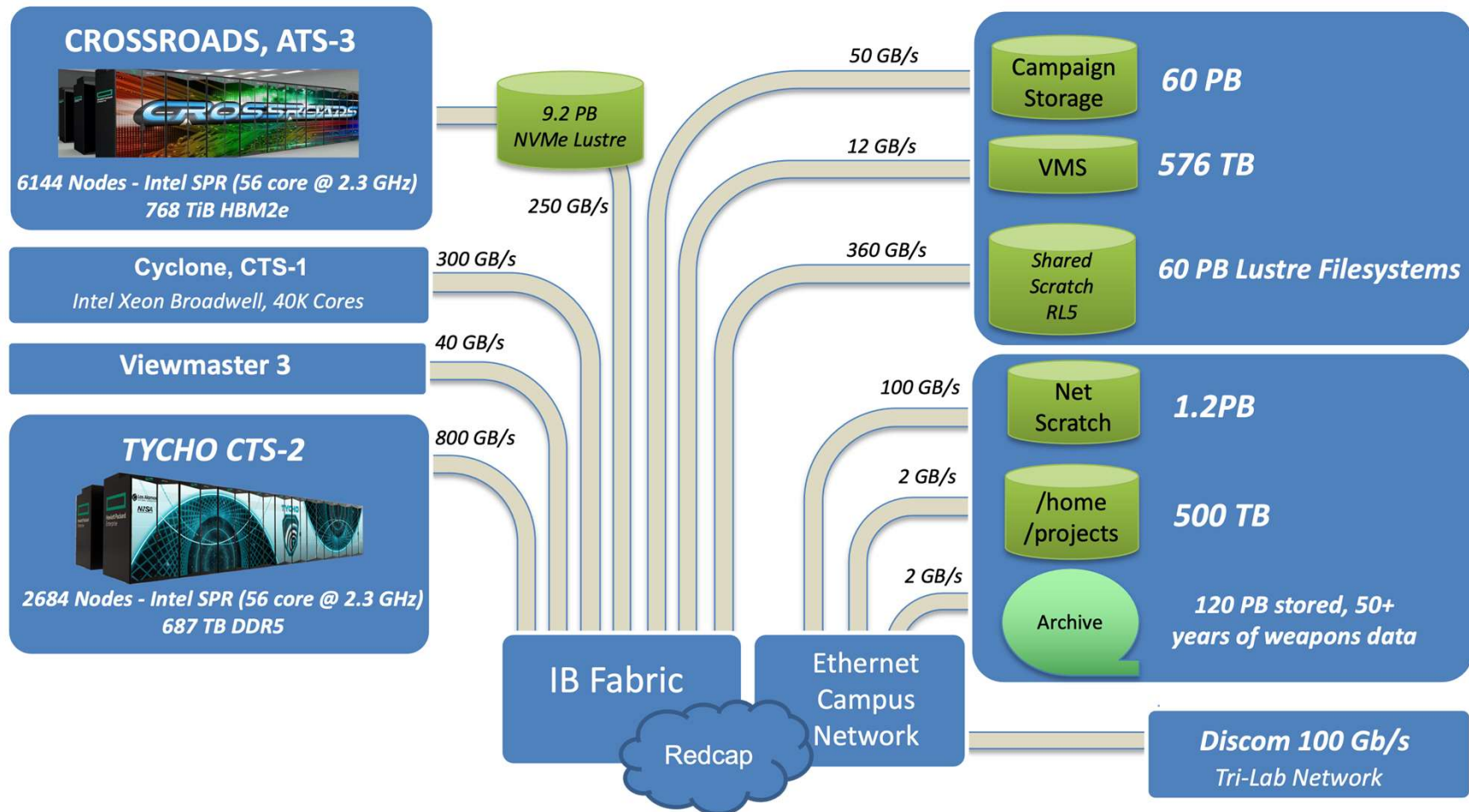


Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

6/4/2026

1

LANL Compute/Storage Environment (Secure) 2025



Sampling of LANL Filesystems in 2025

Filesystem	Directory Count (Millions)	File Count (Millions)
roscratch1	13.9	657.2
users	13.3	93.4
netscratch	1.3	16.8

Grand Unified File Index

<https://github.com/mar-file-system/GUFI>

- Highly parallel for fast index traversal
- Stores metadata in databases
 - Complex queries with SQL
- Enforces permissions so users and admins can use the same index
 - Support for extended attributes
- Single index for all filesystems
- Leverages well developed technologies
 - POSIX filesystem hierarchy, permissions, attributes
 - SQLite 3, PCRE, jemalloc, CMake 3
 - Flash Storage



GRAND UNIFIED FILE INDEX (GUFI)

A Global Meta-Base NOT a Lake House

GUFI enables collecting and querying rich metadata across heterogeneous storage systems and sites, delivering easily consumable results while leaving data in place, respecting source-derived access controls, and reflecting changes over time.

Index Nearly Any Storage System

localFS; NFS; Lustre; HPSS; S3; cloud;
Windows shares;

Extremely Scalable

10-100 Billion items; Billions of folders

Crazy Fast Query

Single seconds for users; Single minutes
for admins

Access Control Derived from Existing Permissions

Reuse of what already exists; ACL/MLS;
POSIX

Support Many Access Methods via SQL

SQL → C/C++, Python, Ruby, R, web, MCP, cmd, beyond...

Multi-Site Access/Deployment

Multiple Different FS; Query between multiple locations

Open Source; Commercial; No Need to Move Data; Award Winning

Available on GitHub; Embedded in 7 Products; 2 Patents;
R&D100 Awarded



HAMMERSPACE

HPE



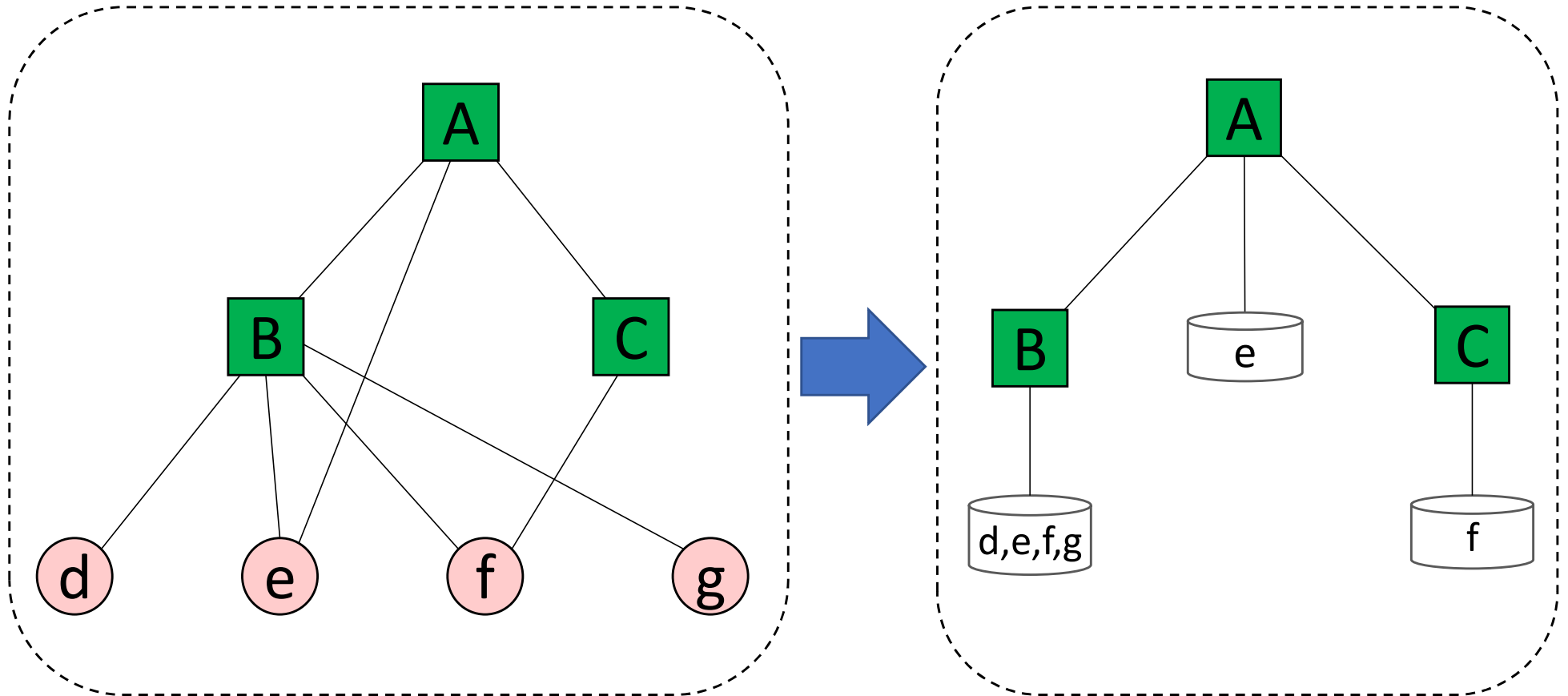
Domino



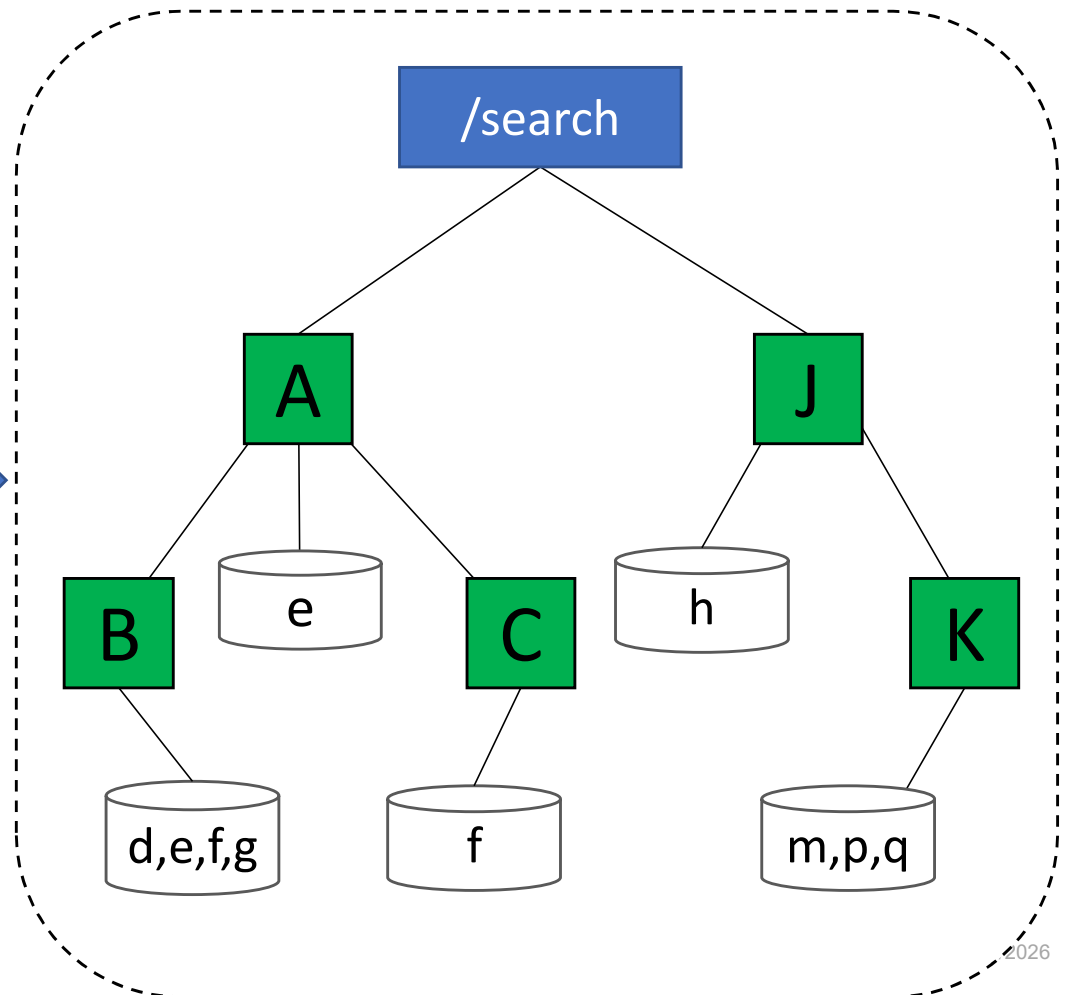
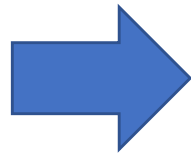
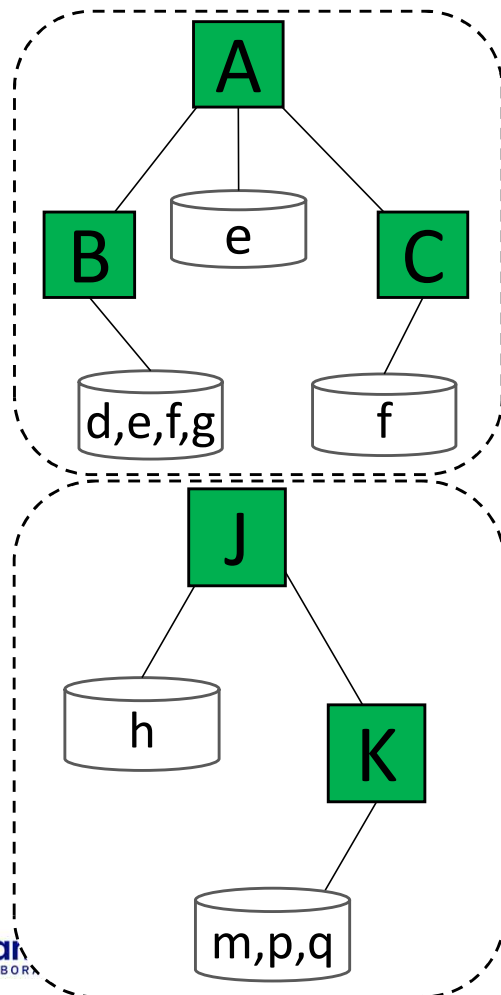
BeeGFS®



Source Filesystem to Index

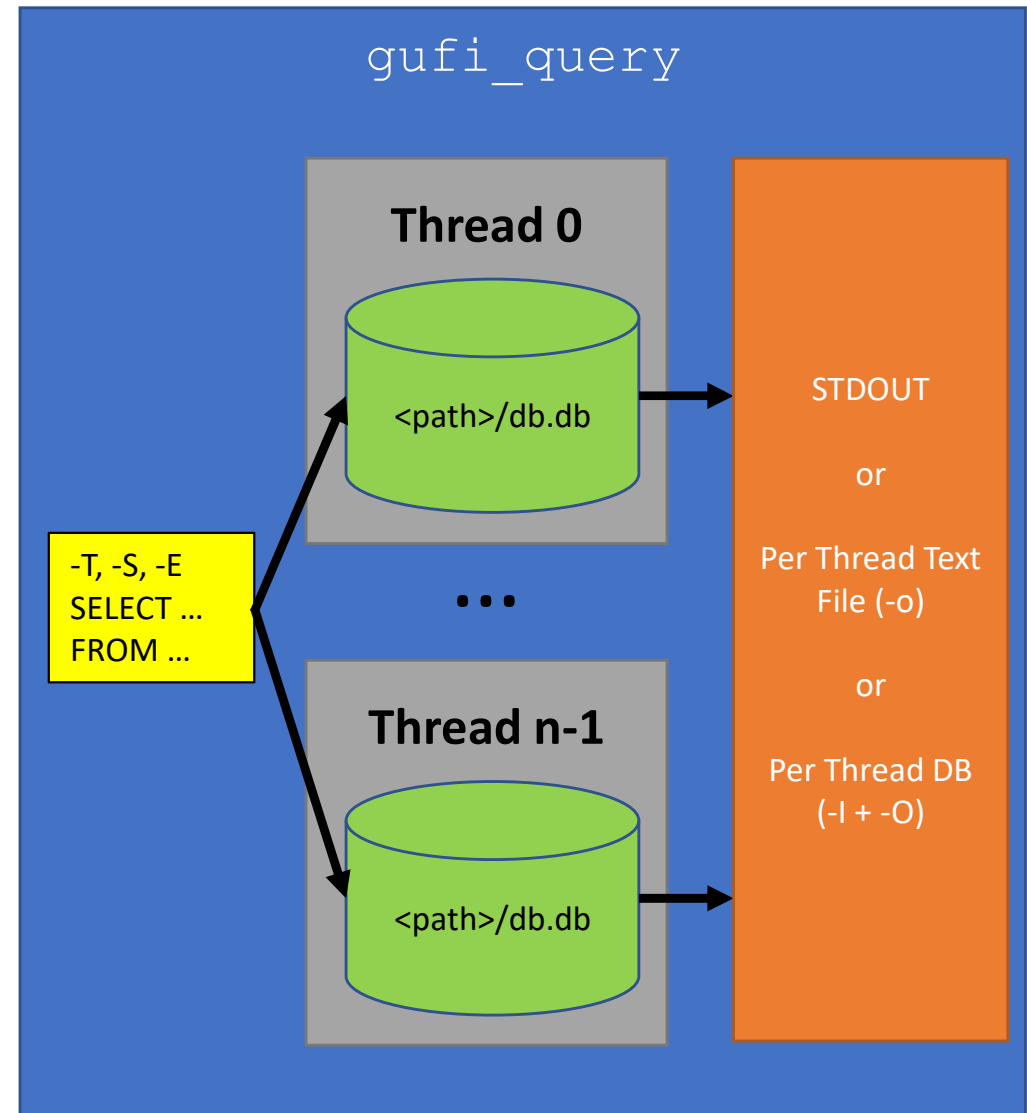


Combining Indices



gufi_query

- Runs raw SQL statements
 - Need to know database and table schemas
 - Meant for admin/advanced users
 - User facing tools wrapping `gufi_query`
- Highly parallel
 - Each directory is processed by a thread



Simple Access Control

```
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
total 16780
-rw-rw-r-- 1 ggrider ggrider ggrider@a30-compute: /tmp/accesstest$
drwxrwx--- 3 ggrider ggrider ggrider@a30-compute: /tmp/accesstest$
-rw-r--r-- 1 ggrider ggrider 1 ggrider@a30-compute: /tmp/accesstest$
root@a30-compute: /tmp/accesstest$ ggrider@a30-compute: /tmp/accesstest$
```

files and directory summaries (allvar) (local) with Traversal Control

Local Traversal Control

Optionally specify traversal control

Search tree: /LOCALSEARCHPATH

tree_minlevel		tree_maxlevel		subtreeowner_uid		subtreeowner_gid	
subtree_search_pathlist_file							
subtree_search_pathlist							

Path Options

Optionally select search path

Search Local Tree

File Input Criteria

Please enter at least one

Path	File Name	Entry Type
number of links	Link Name	xattrname
xattrvalue	File Size	File Size
Entry Type	File UID	File GID
File Mode	File Mod Time	File Mod Time
File Change Time	File Change Time	File Access Time
File Access Time	File Create Time	File Create Time
DLink Name		

files and directory summaries (allvar) (local) with Traversal Control

Local Traversal Control

Optionally specify traversal control

Search tree: /LOCALSEARCHPATH

tree_minlevel tree_maxlevel subtreeowner_uid subtreeowner_gid

subtree_search_pathlist_file

subtree_search_pathlist

Path Options

Optionally select search path

Search Local Tree

File Input Criteria

Please enter at least one

Path	File Name	Entry Type
number of links	Link Name	xattrname
xattrvalue	File Size	File Size
Entry Type	File UID	File GID
File Mode	File Mod Time	File Mod Time
File Change Time	File Change Time	File Access Time
File Access Time	File Create Time	File Create Time
DLink Name		

files and directory summaries (allvar) (local) with Traversal Control

Local Traversal Control

Optionally specify traversal control

Search tree: /LOCALSEARCHPATH

tree_minlevel		tree_maxlevel		subtreeowner_uid		subtreeowner_gid	
subtree_search_pathlist_file							
subtree_search_pathlist							

Path Options

Optionally select search path

Search Local Tree

File Input Criteria

Please enter at least one

Path	File Name	Entry Type
number of links	Link Name	xattrname
xattrvalue	File Size	File Size
Entry Type	File UID	File GID
File Mode	File Mod Time	File Mod Time
File Change Time	File Change Time	File Access Time
File Access Time	File Create Time	File Create Time
DLink Name		

Within Directory Input Criteria

Optionally enter

Directory Name like %str% or 'str1','str2' for in	number of dlinks	Directory UID
Directory GID	Directory Size	Directory Size
Directory Mod Time date input mm/dd/yyyy --:-- --	Directory Mod Time date input mm/dd/yyyy --:-- --	Directory Change Time date input mm/dd/yyyy --:-- --
Directory Change Time date input mm/dd/yyyy --:-- --	Directory Access Time date input mm/dd/yyyy --:-- --	Directory Access Time date input mm/dd/yyyy --:-- --
Min file ctime date input mm/dd/yyyy --:-- --	Max file ctime date input mm/dd/yyyy --:-- --	Min file mtime date input mm/dd/yyyy --:-- --
Max file mtime date input mm/dd/yyyy --:-- --	Min file atime date input mm/dd/yyyy --:-- --	Max file atime date input mm/dd/yyyy --:-- --
Min file ctime date input mm/dd/yyyy --:-- --	Max file crtime date input mm/dd/yyyy --:-- --	Directory depth
Directory dtotxattr	Directory Mode or octal mask	Directory total files
Directory total links	Directory min filesize	Directory max filesize
Directory total zero files	Directory min uid	Directory max uid
Directory min gid	Directory max gid	Directory total files ltk
Directory total files mtk	Directory total files ltm	Directory total files mtm
Directory total files mtg	Directory total files mtt	Directory total files size

Select Fields to retrieve

Please check at least one

☐ Pathname	☐ hostname	☐ linkname	☐ xattrname	☐ xattrvalue	☐ File Size	☐ File Inode	☐ File nlink	☐ Entry Type	☐ File UID	☐ File GID	☐ File Mode	☐ File Mod Time
☐ File Change Time	☐ File Access Time	☐ File Create Time	☐ File DMode	☐ Dir nlink	☐ Dir uid	☐ Dir gid	☐ Dir size	☐ Dir Mod Time	☐ Dir Change Time	☐ Dir Access Time	☐ dlinkname	☐ Dir totfile
☐ Dir totlinks	☐ Dir minuid	☐ Dir maxuid	☐ Dir mingid	☐ Dir maxgid	☐ Dir minsize	☐ Dir maxsize	☐ Dir totzero	☐ Dir totltk	☐ Dir totmtk	☐ Dir totltm	☐ Dir totmtm	☐ Dir totmtg
☐ Dir totmtt	☐ Dir totsize	☐ Dir minctime	☐ Dir maxctime	☐ Dir minmtime	☐ Dir maxmtime	☐ Dir minatime	☐ Dir maxatime	☐ Dir mincrtme	☐ Dir maxcrtme	☐ Dir totxattr		

Order by

Optionally Order by by, you must select a variable in the checkboxes above in order to use it in an order by

First Order By		asc
Second Order By		asc

Within Directory Input Criteria

Optionally enter

Directory Name like %str% or 'str1','str2' for in	number of dlinks	Directory UID
Directory GID	Directory Size	Directory Size
Directory Mod Time date input mm/dd/yyyy --:-- --	Directory Mod Time date input mm/dd/yyyy --:-- --	Directory Change Time date input mm/dd/yyyy --:-- --
Directory Change Time date input mm/dd/yyyy --:-- --	Directory Access Time date input mm/dd/yyyy --:-- --	Directory Access Time date input mm/dd/yyyy --:-- --
Min file ctime date input mm/dd/yyyy --:-- --	Max file ctime date input mm/dd/yyyy --:-- --	Min file mtime date input mm/dd/yyyy --:-- --
Max file mtime date input mm/dd/yyyy --:-- --	Min file atime date input mm/dd/yyyy --:-- --	Max file atime date input mm/dd/yyyy --:-- --
Min file ctime date input mm/dd/yyyy --:-- --	Max file crtime date input mm/dd/yyyy --:-- --	Directory depth
Directory dtotxattr	Directory Mode or octal mask	Directory total files
Directory total links	Directory min filesize	Directory max filesize
Directory total zero files	Directory min uid	Directory max uid
Directory min gid	Directory max gid	Directory total files ltk
Directory total files mtk	Directory total files ltm	Directory total files mtm
Directory total files mtg	Directory total files mtt	Directory total files size

Select Fields to retrieve

Please check at least one

☐ Pathname	☐ hostname	☐ linkname	☐ xattrname	☐ xattrvalue	☐ File Size	☐ File Inode	☐ File nlink	☐ Entry Type	☐ File UID	☐ File GID	☐ File Mode	☐ File Mod Time
☐ File Change Time	☐ File Access Time	☐ File Create Time	☐ File DMode	☐ Dir nlink	☐ Dir uid	☐ Dir gid	☐ Dir size	☐ Dir Mod Time	☐ Dir Change Time	☐ Dir Access Time	☐ dlinkname	☐ Dir totfile
☐ Dir totlinks	☐ Dir minuid	☐ Dir maxuid	☐ Dir mingid	☐ Dir maxgid	☐ Dir minsize	☐ Dir maxsize	☐ Dir totzero	☐ Dir totltk	☐ Dir totmtk	☐ Dir totltm	☐ Dir totmtm	☐ Dir totmtg
☐ Dir totmtt	☐ Dir tosize	☐ Dir minctime	☐ Dir maxctime	☐ Dir minmtime	☐ Dir maxmtime	☐ Dir minatime	☐ Dir maxatime	☐ Dir mincrtme	☐ Dir maxcrtme	☐ Dir totxattr		

Order by

Optionally Order by by, you must select a variable in the checkboxes above in order to use it in an order by

First Order By		asc
Second Order By		asc

Within Directory Input Criteria

Optionally enter

Directory Name like %str% or 'str1','str2' for in	number of dlinks	Directory UID
Directory GID	Directory Size	Directory Size
Directory Mod Time date input mm/dd/yyyy --:-- --	Directory Mod Time date input mm/dd/yyyy --:-- --	Directory Change Time date input mm/dd/yyyy --:-- --
Directory Change Time date input mm/dd/yyyy --:-- --	Directory Access Time date input mm/dd/yyyy --:-- --	Directory Access Time date input mm/dd/yyyy --:-- --
Min file ctime date input mm/dd/yyyy --:-- --	Max file ctime date input mm/dd/yyyy --:-- --	Min file mtime date input mm/dd/yyyy --:-- --
Max file mtime date input mm/dd/yyyy --:-- --	Min file atime date input mm/dd/yyyy --:-- --	Max file atime date input mm/dd/yyyy --:-- --
Min file ctime date input mm/dd/yyyy --:-- --	Max file crtime date input mm/dd/yyyy --:-- --	Directory depth
Directory dtotxattr	Directory Mode or octal mask	Directory total files
Directory total links	Directory min filesize	Directory max filesize
Directory total zero files	Directory min uid	Directory max uid
Directory min gid	Directory max gid	Directory total files ltk
Directory total files mtk	Directory total files ltm	Directory total files mtm
Directory total files mtg	Directory total files mtt	Directory total files size

Select Fields to retrieve

Please check at least one

☐ Pathname	☐ hostname	☐ linkname	☐ xattrname	☐ xattrvalue	☐ File Size	☐ File Inode	☐ File nlink	☐ Entry Type	☐ File UID	☐ File GID	☐ File Mode	☐ File Mod Time
☐ File Change Time	☐ File Access Time	☐ File Create Time	☐ File DMode	☐ Dir nlink	☐ Dir uid	☐ Dir gid	☐ Dir size	☐ Dir Mod Time	☐ Dir Change Time	☐ Dir Access Time	☐ dlinkname	☐ Dir totfile
☐ Dir totlinks	☐ Dir minuid	☐ Dir maxuid	☐ Dir mingid	☐ Dir maxgid	☐ Dir minsize	☐ Dir maxsize	☐ Dir totzero	☐ Dir totltk	☐ Dir totmtk	☐ Dir totltm	☐ Dir totmtm	☐ Dir totmtg
☐ Dir totmtt	☐ Dir totsize	☐ Dir minctime	☐ Dir maxctime	☐ Dir minmtime	☐ Dir maxmtime	☐ Dir minatime	☐ Dir maxatime	☐ Dir mincrtme	☐ Dir maxcrtme	☐ Dir totxattr		

Order by

Optionally Order by by, you must select a variable in the checkboxes above in order to use it in an order by

First Order By		asc
Second Order By		asc

Within Directory Input Criteria

Optionally enter

Directory Name	like >v %str% or "str1","str2" for in	number of dlinks	>v		Directory UID	>v	
Directory GID	>v				Directory Size	>v	
Directory Mod Time	>v date input mm/dd/yyyy --:-- --	Directory Mod Time	>v date input mm/dd/yyyy --:-- --		Directory Change Time	>v date input mm/dd/yyyy --:-- --	
Directory Change Time	>v date input mm/dd/yyyy --:-- --	Directory Access Time	>v date input mm/dd/yyyy --:-- --		Directory Access Time	>v date input mm/dd/yyyy --:-- --	
Min file ctime	>v date input mm/dd/yyyy --:-- --	Max file ctime	>v date input mm/dd/yyyy --:-- --		Min file mtime	>v date input mm/dd/yyyy --:-- --	
Max file mtime	>v date input mm/dd/yyyy --:-- --	Min file atime	>v date input mm/dd/yyyy --:-- --		Max file atime	>v date input mm/dd/yyyy --:-- --	
Min file ctime	>v date input mm/dd/yyyy --:-- --	Max file crtime	>v date input mm/dd/yyyy --:-- --		Directory depth	>v	
Directory dtotxattr	>v	Directory Mode	for >v octal mask		Directory total files	>v	
Directory total links	>v	Directory min filesize	>v		Directory max filesize	>v	
Directory total zero files	>v	Directory min uid	>v		Directory max uid	>v	
Directory min gid	>v	Directory max gid	>v		Directory total files ltk	>v	
Directory total files mtk	>v	Directory total files ltm	>v		Directory total files mtm	>v	
Directory total files mtg	>v	Directory total files mtt	>v		Directory total files size	>v	

Select Fields to retrieve

Please check at least one

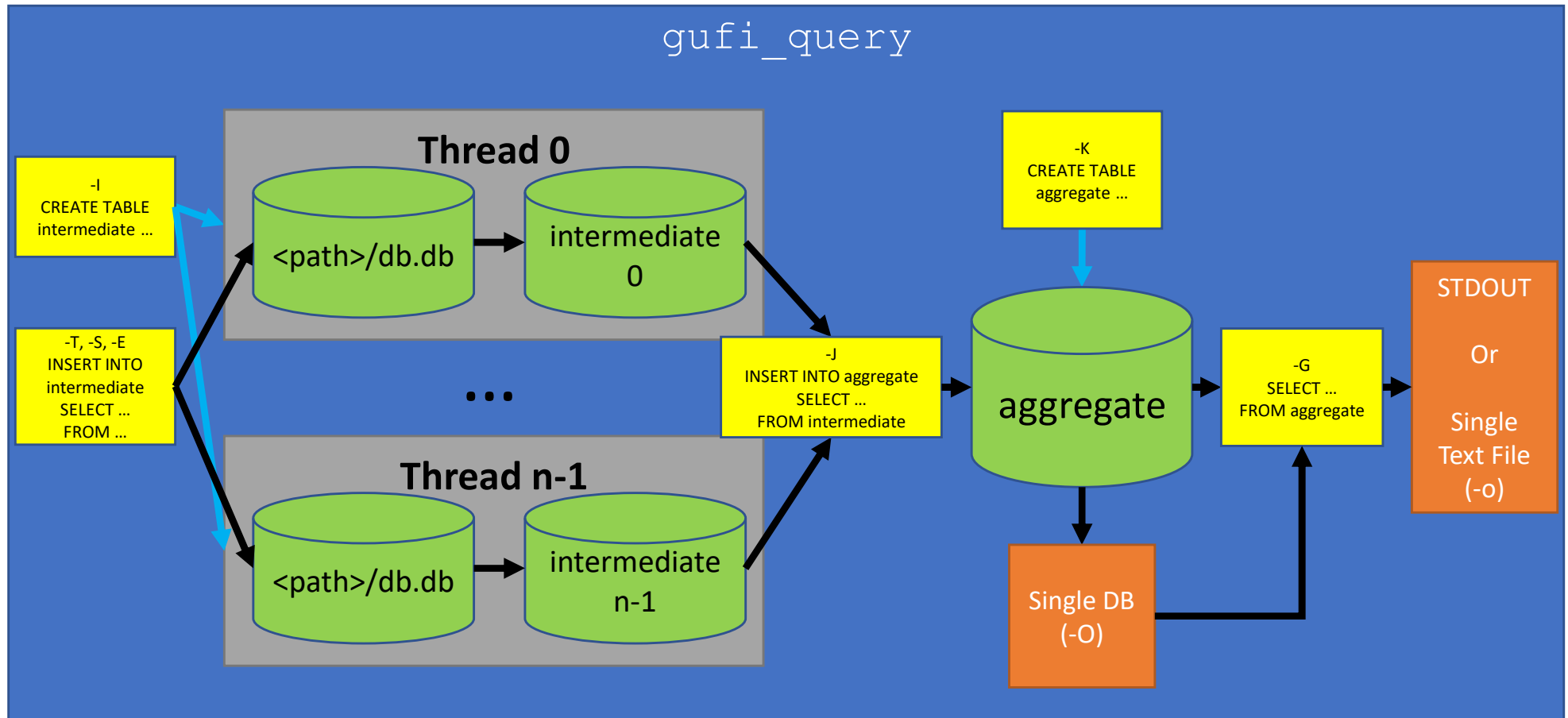
☐ Pathname	☐ hostname	☐ linkname	☐ xattrname	☐ xattrvalue	☐ File Size	☐ File Inode	☐ File nlink	☐ Entry Type	☐ File UID	☐ File GID	☐ File Mode	☐ File Mod Time
☐ File Change Time	☐ File Access Time	☐ File Create Time	☐ File DMode	☐ Dir nlink	☐ Dir uid	☐ Dir gid	☐ Dir size	☐ Dir Mod Time	☐ Dir Change Time	☐ Dir Access Time	☐ dlinkname	☐ Dir totfile
☐ Dir totlinks	☐ Dir minuid	☐ Dir maxuid	☐ Dir mingid	☐ Dir maxgid	☐ Dir minsize	☐ Dir maxsize	☐ Dir totzero	☐ Dir totltk	☐ Dir totmtk	☐ Dir totltm	☐ Dir totmtm	☐ Dir totmtg
☐ Dir totmtt	☐ Dir totsize	☐ Dir minctime	☐ Dir maxctime	☐ Dir minmtime	☐ Dir maxmtime	☐ Dir minatime	☐ Dir maxatime	☐ Dir mincrtme	☐ Dir maxcrtme	☐ Dir totxattr		

Order by

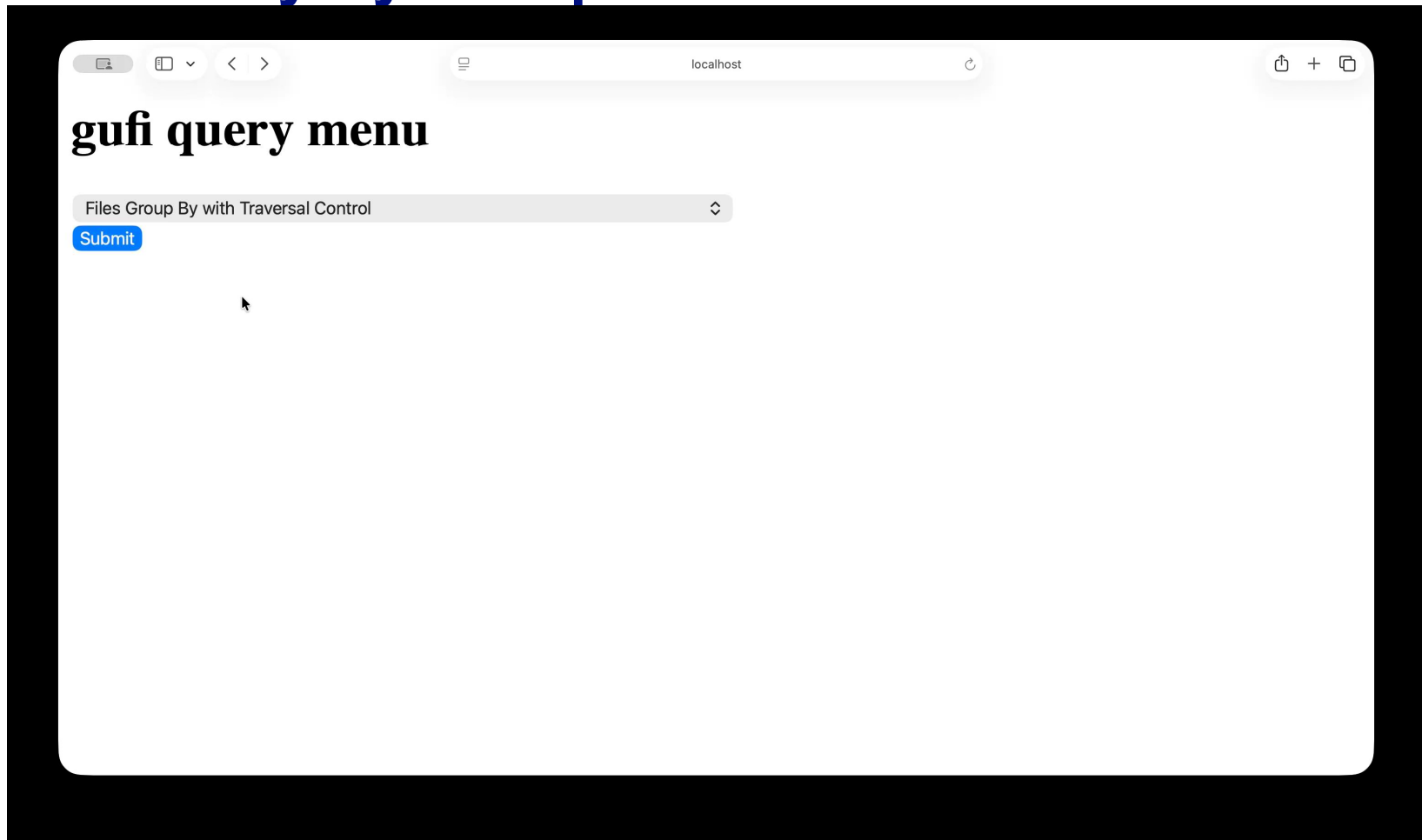
Optionally Order by by, you must select a variable in the checkboxes above in order to use it in an order by

First Order By	>v	asc >v
Second Order By	>v	asc >v

Aggregation



GUFI Query By Group



The screenshot shows a web browser window with the address bar set to 'localhost'. The page title is 'gufi query menu'. Below the title, there is a text input field containing the text 'Files Group By with Traversal Control'. To the right of the input field is a small downward arrow icon. Below the input field is a blue button labeled 'Submit'. The browser window has a dark border and standard navigation buttons (back, forward, home, etc.) in the top left corner.

Virtual Tables

`gufi_vt_*`

- `treesummary`
- `summary`
- `entries`
- `pentries`
- `vrsummary`
- `vrpentries`

`SELECT ...`

`FROM gufi_vt_*(index, threads);`

`gufi_vt`

- Run arbitrary queries

```
CREATE VIRTUAL TABLE
temp.gufi
USING gufi_vt(index,
threads=2, S="...", E="...");
```

`SELECT ...`

`FROM temp.gufi;`

Use gufi_vt Anywhere

- Python

- SQLite 3

- Use numpy, pandas, etc. to analyze data

- SQLAlchemy

- The Python SQL Toolkit and Object Relational Mapper
 - <https://www.sqlalchemy.org/>
 - Now can use SQLAlchemy frontends with GUF

```
with sqlite3.connect(":memory:") as conn:
    conn.enable_load_extension(True)
    conn.load_extension(gufi_vt)
```

- RSQLite

```
conn <- dbConnect(RSQLite::SQLite(), ":memory:", loadable.extensions = TRUE)
dbExecute(conn, "SELECT load_extension('gufi_vt');")
```

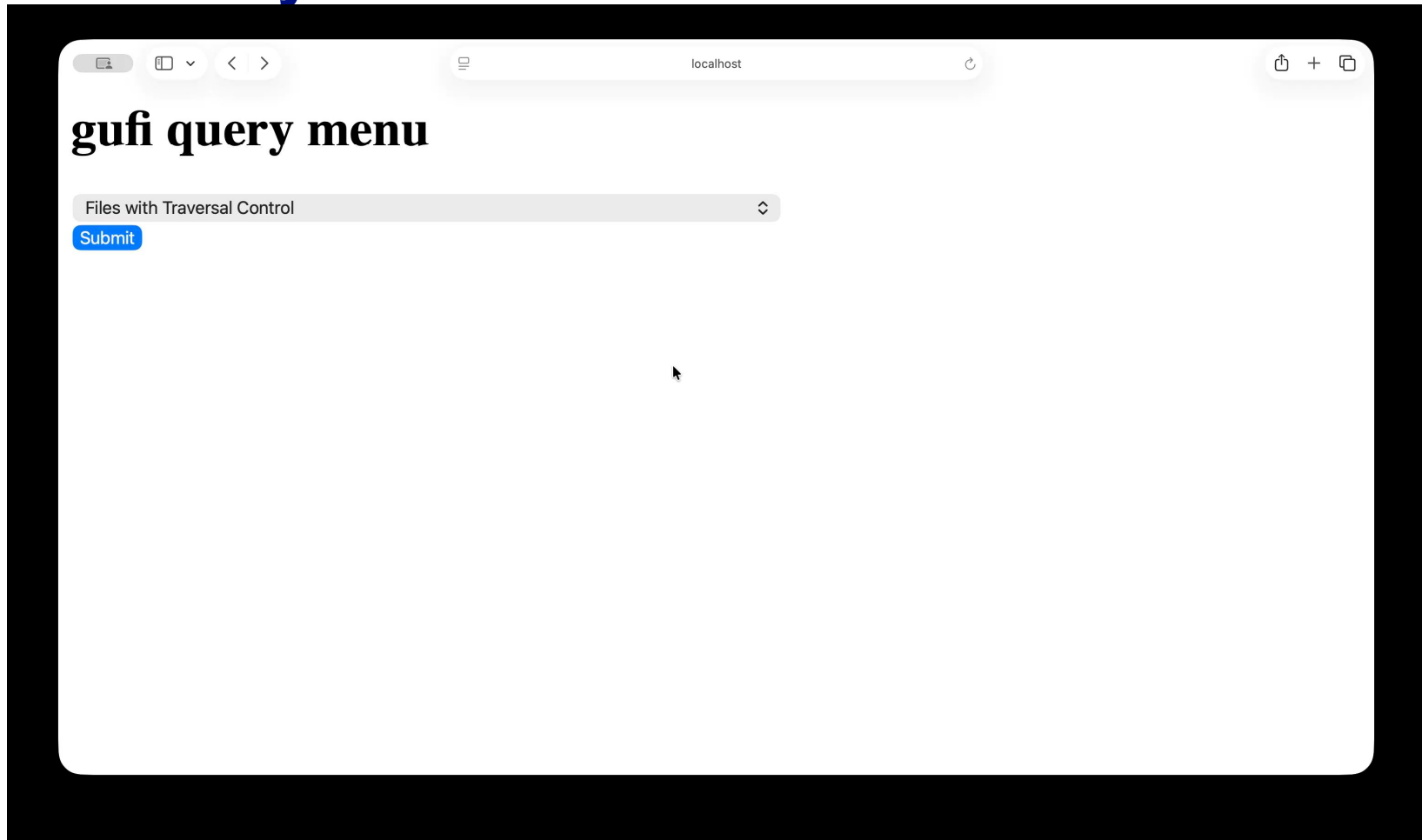
- sqlite3-ruby

```
conn = SQLite3::Database.new(':memory:')
conn.enable_load_extension(true)
conn.load_extension("gufi_vt")
```

Querying Remotes

- `gufi_vt + ssh`
 - Pull from single remote
 - Each index may be at a different path on each remote
- `gufi_vt + pdsh`
 - Pull from all remotes in parallel
 - All indexes must be at the same path on each remote
- Example implementation: `gufi_query_multiple_remotes`

GUFI Query Over Two Sites



The screenshot shows a web browser window with the address bar set to 'localhost'. The page title is 'gufi query menu'. Below the title, there is a text input field containing the text 'Files with Traversal Control'. To the right of the input field is a small downward arrow icon. Below the input field is a blue button labeled 'Submit'.

More Information

- Source Code, Documentation, and Previous Presentations
 - <https://github.com/mar-file-system/GUFI>
 - Contributions are welcome!
- Dominic Manno, Jason Lee, Prajwal Challa, Qing Zheng, David Bonnie, Gary Grider, and Bradley Settlemyer. 2022. GUFI: fast, secure file system metadata search for both privileged and unprivileged users. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '22). IEEE Press, Article 57, 1-14.

Access Controlled AI Context

External Databases

- Allows for arbitrary extra information to be placed alongside metadata
- User generated
- System generated

Store Extracted Text in Index

- SQLite 3 FTS5
 - Natural Language Processing
- PDF Metadata
- Image Metadata + OCR
- Extensible
 - netcdf, hdf5, vtk
- VaultIQ
 - Microsoft Office
 - Image feature extraction
 - tarball/zip
 - Code

Base + Text & Feature Extraction (Search Demo)

Please enter at least one

File Name like File Size >

File Mod Time > wordf match

Select Fields to retrieve

Please check at least one

☒ Pathname ☒ hostname ☒ File Size ☐ File Mod Time ☒ Words ☒ Rank

Order by

Optionally Order by by

First Order By

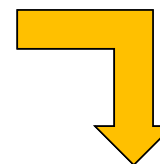
Second Order By

verbose_level

Base + Text & Feature Extraction (Closer Look)

Test System with 70k Files; 7k Directories

File Name like File Size >
File Mod Time > wordf match



pathname	hostname	size	wordf	rank
/search/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/CCN-9/ISTI/Cluster-Network-Summer-School-FY16/ISTI-Cluster-Network-Summer-SchoolHPC Overview-2016.pptx	regufi.lanl.gov	21639992	...isti summer school director -->amanda<-- -- >bonnie<-- nmc management administration andree...	-0.794096830488506
/search/historyidx/Z-GaryGriderHistory/recent-presentations/ISTI-Cluster-Network-Summer-SchoolHPC Overview-2016.pptx	regufi.lanl.gov	21649859	...isti summer school director -->amanda<-- -- >bonnie<-- nmc management administration andree...	-4.89971868308536

total rows 2

Base + Text & Feature Extraction (more)

Looking at the database content we list:

File Name, Asset Type, Extracted File Metadata and Text

Note the “derived” and having one or more records for a given file

```
Roadrunner System Overview (LA-UR-08-2087).ppt|derived|0|{"title":  
"Roadrunner Project Preview", "author": "ESD Registered User",  
"last_modified_b|  
|Document Metadata for Roadrunner System Overview (LA-UR-08-  
2087).ppt Title: Roadrunner Project Previ  
Roadrunner System Overview (LA-UR-08-2087).ppt|derived|1|{"title":  
"Roadrunner Project Preview", "author": "ESD Registered User",  
"last_modified_b|
```

Base + Text & Feature Extraction (more)

Notice how it pulled specific Microsoft metadata from the file

```
ggrider_hpcwire.pptx|derived|0|{"title": "Godfather of Parallel  
File Systems", "author": "Grider, Gary", "keywords": "hpc|  
|Document Metadata for ggrider_hpcwire.pptx Title: Godfather or  
Parallel File Systems Author: Grider,
```

This was a picture with no text or the model didn't have anything to say about it
(the actual image was a small logo)

```
ggrider_hpcwire.pptx|derived|1|{"title": "Godfather of Parallel  
File Systems", "author": "Grider, Gary", "keywords": "hpc|  
|Image /page/1/Picture/0
```

Base + Text & Feature Extraction (Nested)

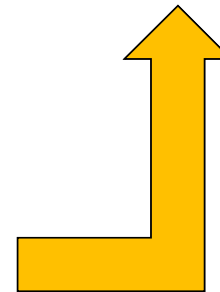
It found a picture on page 2 and performed OCR on it

```
ggrider_hpcwire.pptx|derived|2|{"title": "Godfather of Parallel  
File Systems", "author": "Grider, Gary", "keywords": "hpc|  
|35 HPC Legends Gary Grider Gary Grider Godfather of Parallel  
File Systems Image /page/2/Picture/3
```



Gary Grider

Godfather of Parallel File Systems



Vector Databases

- Generate embeddings
 - <https://github.com/rodydavis/sqlite-lembed>
 - Fork of <https://github.com/asg017/sqlite-lembed>
 - .gguf format
 - llama.cpp
 - Local LLM Inference
- Store vector embeddings and calculate vector distances
 - <https://github.com/asg017/sqlite-vec>

Timing a RAG Query

[illegible]

Closer Look at RAG Query

```
ggrider@pn2201328 GUFi-MAC-Indexer %  
ggrider@pn2201328 GUFi-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null ]  
  
**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided ****  
  
***** top N files *****  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG  
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG  
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx  
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...  
  
**** context *****  
roadrunner andy white anastasio scc pete dominici pp p  
pete dominici andy white roadrunner scc fastest earmark anastasio pp p  
computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc  
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur  
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil  
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastru  
cture capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e  
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity  
acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site  
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi  
  
*****  
  
question: who was andy white  
NO CONTEXT  
  
WITH GUFi RAG CONTEXT  
and what was his role in the development of the fastest computer in the world?  
Answer: andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was  
responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa  
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a  
thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.  
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 8.29s user 1.97s system 115% cpu 8.865 total  
ggrider@pn2201328 GUFi-MAC-Indexer %
```


Closer Look at RAG Query



```
ggrider@pn2201328 GUFi-MAC-Indexer %  
ggrider@pn2201328 GUFi-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null ]
```

**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided ****

***** top N files *****

```
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG  
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG  
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx  
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...
```

**** context ****

```
roadrunner andy white anastasio scc pete dominici pp p  
pete dominici andy white roadrunner scc fastest earmark anastasio pp p  
computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc  
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur  
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil  
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastru  
cture capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e  
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity  
acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site  
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi
```

```
question: who was andy white  
NO CONTEXT
```

WITH GUFi RAG CONTEXT

```
and what was his role in the development of the fastest computer in the world?  
Answer:andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was  
responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa  
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a  
thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.  
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 8.29s user 1.97s system 115% cpu 8.865 total  
ggrider@pn2201328 GUFi-MAC-Indexer %
```


Closer Look at RAG Query

```
ggrider@pn2201328 GUFi-MAC-Indexer %  
ggrider@pn2201328 GUFi-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null ]
```

**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided ****

***** top N files *****

file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...

**** context ****

roadrunner andy white anastasio scc pete dominici pp p
pete dominici andy white roadrunner scc fastest earmark anastasio pp p
computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastru
cture capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity
acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi

question: who was andy white
NO CONTEXT

WITH GUFi RAG CONTEXT

and what was his role in the development of the fastest computer in the world?
Answer: andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was
responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a
thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 8.29s user 1.97s system 115% cpu 8.865 total
ggrider@pn2201328 GUFi-MAC-Indexer %

Closer Look at RAG Query

```
ggrider@pn2201328 GUFi-MAC-Indexer %  
ggrider@pn2201328 GUFi-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null ]  
  
**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided ****  
  
***** top N files *****  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG  
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG  
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx  
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...  
  
**** context *****  
roadrunner andy white anastasio scc pete dominici pp p  
pete dominici andy white roadrunner scc fastest earmark anastasio pp p  
computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc  
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur  
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil  
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastru  
cture capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e  
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity  
acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site  
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi  
  
*****  
  
question: who was andy white  
NO CONTEXT  
  
WITH GUFi RAG CONTEXT  
and what was his role in the development of the fastest computer in the world?  
Answer: andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was  
responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa  
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a  
thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.  
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 8.29s user 1.97s system 115% cpu 8.865 total  
ggrider@pn2201328 GUFi-MAC-Indexer %
```

Closer Look at RAG Query

```
ggrider@pn2201328 GUFi-MAC-Indexer %  
ggrider@pn2201328 GUFi-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null ]
```

**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided ****

***** top N files *****

```
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG  
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG  
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx  
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...
```

**** context *****

```
roadrunner andy white anastasio scc pete dominici pp p  
pete dominici andy white roadrunner scc fastest earmark anastasio pp p  
computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc  
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur  
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil  
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastru  
cture capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e  
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity  
acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site  
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi
```

question: who was andy white
NO CONTEXT

WITH GUFi RAG CONTEXT

```
and what was his role in the development of the fastest computer in the world?  
Answer:andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was  
responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa  
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a  
thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.  
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 8.29s user 1.97s system 115% cpu 8.865 total  
ggrider@pn2201328 GUFi-MAC-Indexer %
```


Closer Look at RAG Query

```
ggrider@pn2201328 GUFi-MAC-Indexer %  
ggrider@pn2201328 GUFi-MAC-Indexer % time ./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 2>/dev/null ]
```

**** text from gufi using input keyword (full text search) and vector search nearest neighbor to question provided ****

***** top N files *****

```
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4025.JPG  
snippet: roadrunner -->andy<-- -->white<-- anastasio scc pete dominici pp p  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/pictures/evenmorepics/_MG_4023.JPG  
snippet: pete dominici -->andy<-- -->white<-- roadrunner scc fastest earmark anastasio pp...  
file: /Users/ggrider/vaultiq/historyidx/Z-GaryGriderHistory/older-presentations/106273ppt/HPC-DO/lanl_computing.pptx  
snippet: ...role -->andy<-- -->white<-- marie lead lufd kurt soo -->andy<-- duncan...
```

**** context ****

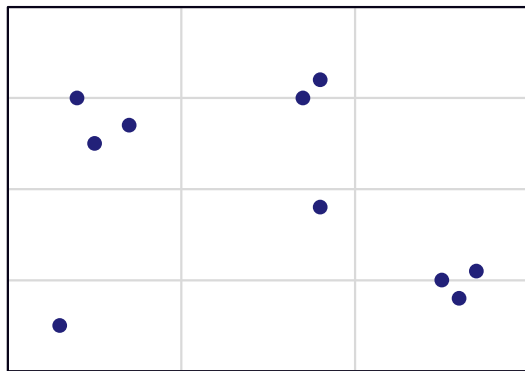
```
roadrunner andy white anastasio scc pete dominici pp p  
pete dominici andy white roadrunner scc fastest earmark anastasio pp p  
computing lanl vision future april goal computing recognized leader high performance computing especially national security scienc  
e forefront computer science key elements laboratory computing program scope asc institutional computing sci computing key signatur  
es success visible integrated leadership compelling documented strategy robust institutional investment xcommitment including facil  
ity plan scope asc computing successful delivery trinity beyond excellent support mission stable reliable computational infrastru  
cture capability capacity machines responsive customer service leadership middleware computing paradigms effective partner llnl snl e  
specially aces asc effective exascale partnership scope institutional computing sustained commitment multi year spend plan capacity  
acquisition best class ability explore novel architectures test beds data intensive virtual clusters including partners xoff site  
strategic allocation resources tie winning proposals transparent reliable service users responsive trends bi
```

```
question: who was andy white  
NO CONTEXT
```

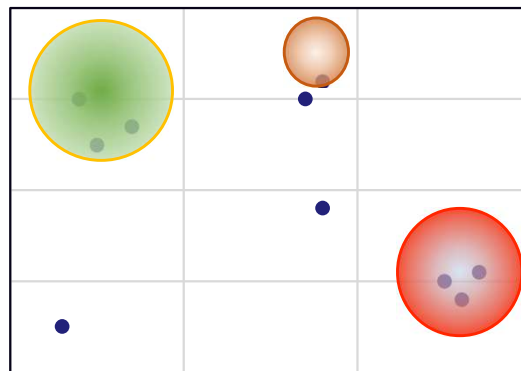
WITH GUFi RAG CONTEXT

```
and what was his role in the development of the fastest computer in the world?  
Answer: andy white was the chief scientist of the supercomputer center (scc) at the national laboratory (lanl) in new mexico. He was  
responsible for the design and development of the world's fastest computer, which was christened the roadrunner. The roadrunner wa  
s a 100-petaflop supercomputer that was built in 1998. It was the first supercomputer to achieve a petaflop performance, which is a  
thousand trillion floating-point operations per second. White was also the director of the supercomputing center at the llnl.  
./run_history_vec_conf 'andy white' 'andy AND white' 'who was andy white' '' 8.29s user 1.97s system 115% cpu 8.865 total  
ggrider@pn2201328 GUFi-MAC-Indexer %
```

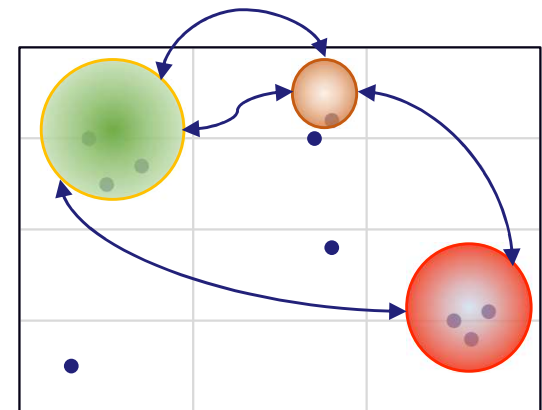
Next Steps



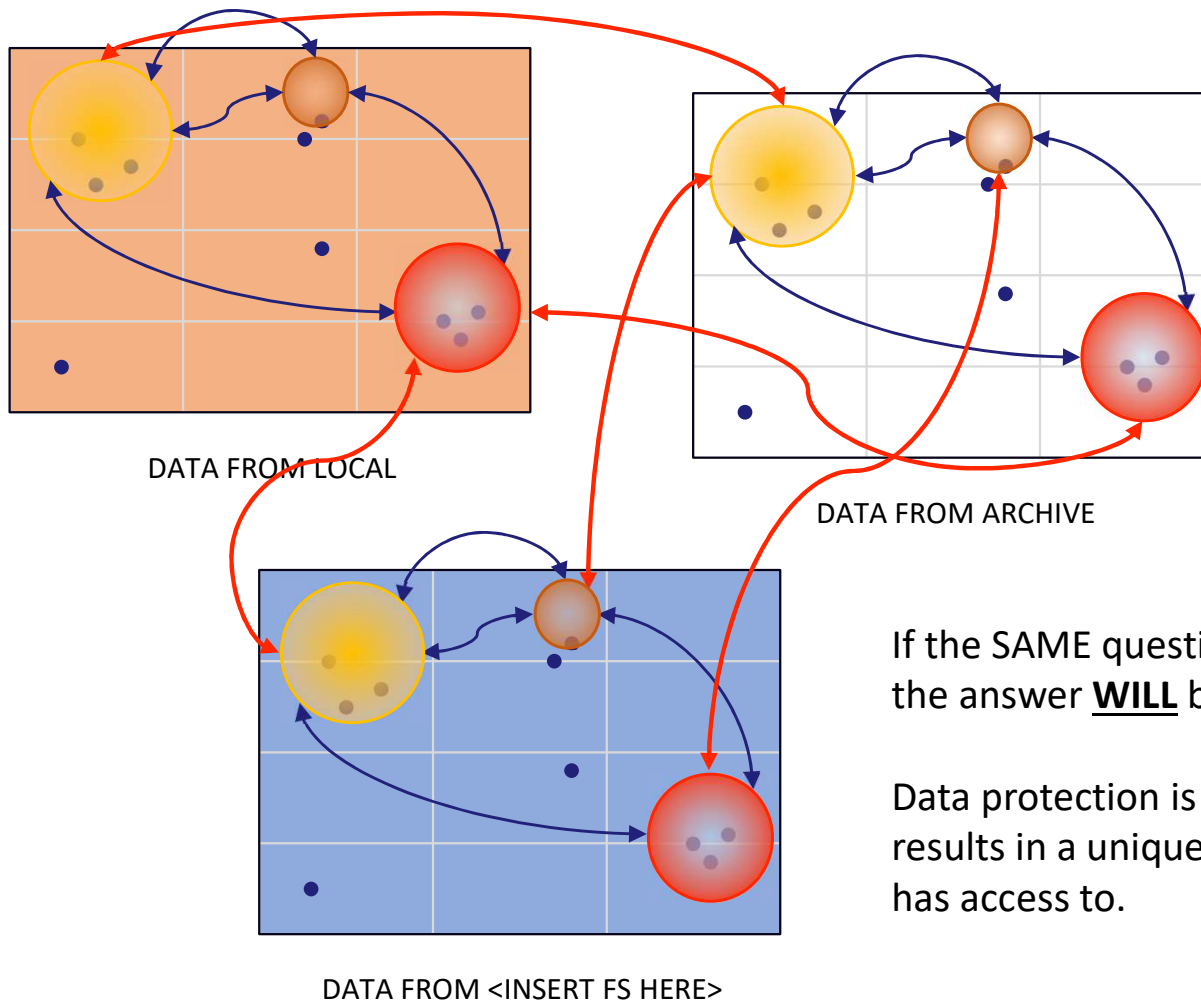
GET (MORE) VECTORS



FIND NEAREST NEIGHBORS
& CLASSIFY NEIGHBORS



FIND NEIGHBOR CONNECTIONS

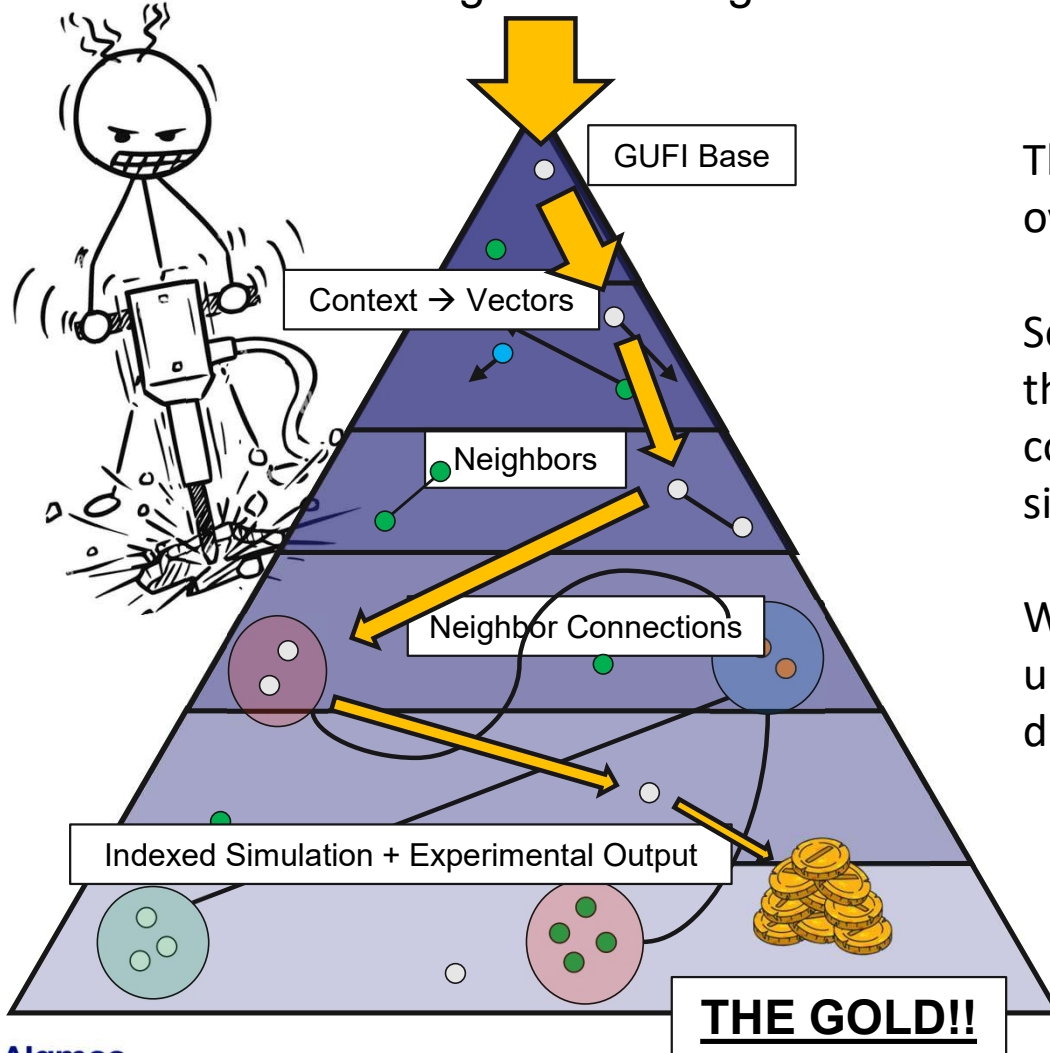


Relationships can be made to different sets of data that one has access to.

If the SAME question is asked by a *different* person, the answer **WILL** be different.

Data protection is maintained throughout; this results in a unique answer based on what data one has access to.

Providing a Narrowing Route to the GOLD



The real value lies hidden within overwhelming data—

So instead of relying solely on SMEs that originally collected the data or conducted the simulation/experiment...

We need tools like GUFi to uncover unseen connections and rapidly drill down to the critical insights.

Thank you!

Database Schema

Summary Table

Dir Name	Proj1
Dir Inode Num	23
Dir UID	7
Dir GID	0
Total Files	3
Min-Max UID	0-7
Min-Max GID	0-1

Entries Table

File Name	Inode Num	UID	GID	Mode	...	PInode
a.out	624	0	0	644		23
main.cc	56	7	0	644		23
1.log	334	2	1	400		23

Pentries View

User Facing Tools

Standard Tool	GUFi Equivalent
du	gufi_du
getfattr	gufi_getfattr
find	gufi_find
ls	gufi_ls
stat	gufi_stat

Base + Text & Feature Extraction (more)

This is just extracted text from the slide

```
ggrider_hpcwire.pptx|derived|3|{"title": "Godfather of Parallel  
File Systems", "author": "Grider, Gary", "keywords": "hpc|  
|Gary Grider is Senior Director for Computing Technologies at  
Los Alamos National Laboratory where he
```