ClarityNow Architecture Guide

Abstract

ClarityNow is an optimized data storage scan, index, and in-memory search database platform that provides visibility, without requiring data access, to data spanning multiple platforms. Flexible plugins for functions like data transfer and audited delete allow data asset stakeholders to quickly manage data resources across primary storage and archive platforms such as Dell EMC® Isilon™ scale-out Network Attached Storage (NAS) and Dell EMC® ECS™ object storage.

January 2019
Revisions

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>December 2018</td>
<td>Initial release</td>
</tr>
</tbody>
</table>

Acknowledgements

This paper was produced by the following members of the Dell EMC storage engineering team:

Author: Lynn Ragan
Support:
Other:

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

© h17521.2 January 2019 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, Dell EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.
# Table of contents

**Acknowledgements** .......................................................................................................................................................... 2
**Revisions** ............................................................................................................................................................................. 2

**Executive Summary** ......................................................................................................................................................... 5
**Introduction** ......................................................................................................................................................................... 6
  - Audience ........................................................................................................................................................................... 6
  - Scope ............................................................................................................................................................................... 6

**Business challenges of unstructured data** ....................................................................................................................... 7
**ClarityNow Solution Overview** ........................................................................................................................................ 8
**ClarityNow use cases and solution positioning** .................................................................................................................. 9
**Key functional benefits and features** ............................................................................................................................. 10
  - Assign TCO costs and view storage consumption costs ........................................................................................................ 10
  - Management level view of data storage without access .................................................................................................... 11
  - Out of band file indexing and data movement ................................................................................................................... 11
  - Cross-platform functionality ............................................................................................................................................. 12
  - Reporting .......................................................................................................................................................................... 12
  - Individual Folder/File tracking ........................................................................................................................................ 13

**ClarityNow Rich Feature/Functionality** ............................................................................................................................ 14
  - File indexing via rapid filesystem scan ................................................................................................................................... 14
  - Auto-tagging ...................................................................................................................................................................... 15
  - Auto-Tag report example ................................................................................................................................................... 16
  - Rule based tagging options ................................................................................................................................................ 19
  - Folder tree fast search with index DB ................................................................................................................................. 21
  - Tactical file and folder movement to archive targets ........................................................................................................ 22
  - Delegated user permissions ................................................................................................................................................ 22
  - LDAP integration ................................................................................................................................................................. 24
  - Cross platform / Multiplatform support ............................................................................................................................ 24
  - Action cart ......................................................................................................................................................................... 25
  - Offline scan log analysis ...................................................................................................................................................... 25

**Architectural** ....................................................................................................................................................................... 27
  - ClarityNow Server ............................................................................................................................................................. 27
  - Server IndexDB and SSD .................................................................................................................................................... 28
  - Multiple communication channels to ClarityNow server ................................................................................................ 29
  - ClarityNow client and web reports access ........................................................................................................................ 29
  - Data Mover ....................................................................................................................................................................... 31
  - Capability of data mover units ........................................................................................................................................... 31
Acknowledgements

Distribution of components on server host and on data mover host.................................................................32
Data Mover service components........................................................................................................................32
Data mover plugin overall capability..................................................................................................................34
Anatomy of data move process ............................................................................................................................35
Choosing a file or folder ....................................................................................................................................36
Transfer to target..................................................................................................................................................37
Validate data move...............................................................................................................................................37
Conclusion ............................................................................................................................................................39
Executive Summary

Data assets in the form of unstructured data continue to challenge companies who struggle to manage their explosive storage growth requirements. Additionally, managing the costs of storage growth and maintaining efficiently organized data paths is especially difficult for the media & entertainment, life sciences, biotechnologies, and financial industries. These types of workflows usually contain large amounts of data that are frequently copied, moved, and retained, usually through a manual process. Decisions to move data to other folder structures, or other storage platforms are also manual and made at within tactical, non-cyclical, timeframes.

Common storage issues for these types of businesses include:

- Difficulty finding digital assets in a timely manner in deep directory structures over distributed storage platforms
- Storage platforms filling to capacity while being unorganized and lacking context causes difficulty for storage administrators to discern what data can be safely archived or deleted
- High storage administration overhead incurred while tracking the amount of storage used by individual customers, employees, departments, projects, and programs.
- Restricting access to digital assets on a need-to-know basis, without requiring unauthorized personnel to contact and disrupt authorized personnel for non-sensitive information related to those assets.

Dell EMC presents ClarityNow as a single pane file/object management solution for complex unstructured environments. ClarityNow is an advanced filesystem scan, index, and fast search platform that provides visibility without access to critical data assets. Flexible plugins for data transfer, audited delete, and many other functions, comprise an immediate take-action toolset for data asset stakeholders seeking to manage data resources across primary storage and archive platforms such as Dell EMC® Isilon™ scale out Network Attached Storage (NAS) and Dell EMC® ECS™ native object archive (S3).

A ClarityNow server quickly scans and indexes volumes provide by network available NFS, SMB or S3 storage platforms. At the same time, it provides a way to customize the classification and tagging of data according to business context as determined by the data asset stakeholder. Fast Search capabilities enable data administrators to quickly find and organize individual files and folders and generate report summaries based on business context autotags. All of this information is made visible without granting access to the actual data content.

Finally, ClarityNow provides a plugin ready API to enable on the spot management decisions about the data to be enacted. The data mover plugin allows for a more surgical approach to data copy or move (copy with delete) of specifically selected files or folders from network filesystems to S3 targets (either cloud based or on premise). The data mover plugin contrasts with other bulk data migration tools such as Isilon Cloudpools in that it is not an automated time-based data-stubbing engine which handles data in terms of large scope policy engines. All data movement is out-of-band in relation to normal file and object access.
Introduction

ClarityNow provides a unique way to manage unstructured data stored across multiple platforms including file and object, on either local or cloud based data platforms. ClarityNow is an advanced filesystem scan, index, and fast search platform that provides visibility without access to critical data assets. Flexible plugins for data transfer, audited delete, and many other functions, comprise an immediate take-action toolset for data asset stakeholders seeking to manage data resources across primary storage and archive platforms such as Dell EMC® Isilon™ scale out Network Attached Storage (NAS) and Dell EMC® ECS™ native object archive (S3).

Audience

This content is intended to enable the presales conversation between technical sales and customer about a ClarityNow solution for cross platform indexing and reporting on unstructured data. Customers may find this paper useful for conceptualizing application of ClarityNow to their particular workflows and business use cases. This paper also raises key points of consideration, preferred use cases, and deployment architecture for presales who are positioning ClarityNow as an unstructured data indexing and management solution.

Scope

This paper addresses the primary uses cases of ClarityNow and its primary feature functionality. ClarityNow business value proposition is outlined and compared with known business challenges. Additionally, the general architecture of the software solution is broken down with detail provided about each functional aspect of deployment.

Day-to-day systems administration, security administration, customized API usage, and POC planning are out of scope for this paper.
Business challenges of unstructured data

Businesses accumulate data assets at an ever accelerating pace. Workable content gets stored either as files or objects, on locally dispersed filesystem platforms, or object stores which are either on premise or cloud-based subscription services. As these repositories grow, and as the business grows, the line between active data and archived data becomes clouded and difficult to visualize. This presents several business challenges in terms of personnel resources, storage costs, business workflow efficiency and overall security.

Most businesses can summarize these problems down to a set of practical questions which might include:

- Where is my data across all my platforms?
- How much does it cost for my data to be stored where it is?
- How can I see which files are aging and no longer needed in production?
- Is there any way I can categorize my data so I can make case by case decisions about where to store it?
- Can I see timely reports which allow me to focus on specific project or team files and folders to determine further retention actions?
- How can I provide an authorized view of my data storage, without providing access to my actual data content?
- How can I enable my department/data managers to find what they need on their own?

Customer infrastructure and data asset managers face questions like these daily. These business challenges are common across several industries and can have significant repercussions in terms of staff time/cost utilization, inefficient cost storage over time at a macro level, and inability to see how costs are applied to specific business classification of data.

Most businesses ultimately want to be able to find their data assets quickly, categorized in a way that makes sense to their unique business workflow. Additionally, as data grows, it becomes critical to index and manage archiving locations (either local or remote) in order to stay paced with ever changing storage platform options. At the same time, as the sheer size of data archive content expands, the challenge increases to provide appropriately controlled access in terms of data management views without allowing indiscriminate access to data files.
ClarityNow Solution Overview

ClarityNow provides a unique way to manage unstructured data stored across multiple platforms including file and object, local and cloud based. Solution capabilities include

- Near real-time scan of all unstructured data
- Analysis and ‘tagging’ of business and project data
- High-speed search across heterogeneous file systems and storages
- Report on the true cost of dormant and redundant data
- Generate showback / chargeback views
- Migrate and move data to archive (i.e. ECS) and cloud tiers

![ClarityNow application overview](image)

Figure 1  ClarityNow application overview

A ClarityNow server quickly scans and indexes volumes provide by network available NFS, SMB or S3 storage platforms. At the same time, it provides a way to customize the classification and tagging of data according to business context as determined by the data asset stakeholder. Fast Search capabilities enable data administrators to quickly find and organize individual files and folders and generate report summaries based on business context autotags. All of this information is made visible without granting access to the actual data content.

Finally, ClarityNow provides a plugin ready API which enables management decisions about the data such as storage platform location, business context tags, audited file deletion, and many others. The data mover plugin allows for a more surgical approach to data copy or move (copy with delete) of specifically selected files or folders from network filesystems to S3 targets (either cloud based or on premise). The data mover plugin contrasts with other bulk data migration tools such as Isilon Cloudpools in that it is not an automated time-based data-stubbing engine which handles data in terms of large scope policy engines. All data movement is out-of-band in relation to normal file and object access.
ClarityNow Solution

ClarityNow use cases and solution positioning

The data workflows addressed by ClarityNow have some common themes. Likely use cases focus where file content is organized around specific business team structures or business processes in which the context is critical to specific data asset managers or stakeholders, and where decisions about the data are handled on a case by case basis. In addition, file data becomes eligible for archive at irregular intervals according to project, or data usage, requirements. Finally, when data is deemed ready for archive, it must be moved out-of-band, while not relying on any stubbing technology for search and retrieval. Data files must be able to be individually restored to production platforms upon demand.

Because of these needs and requirements, ClarityNow is best positioned for data workflows that need human management and decision structure rather than policy based automation or regularly scheduled auto-tiering jobs such as is found with Dell EMC Isilon native Cloudpools tiering technology. Claritynow provides an alternative method of data tiering between Isilon and ECS, providing tight integration between the two storage platforms, while offering a more surgical approach to data management tiering decisions.

Data administrators and managers become the critical stakeholders for this solution which is designed to meet their daily file organizational needs, provide fast search and data context reporting, and optional plugin enabled data movement, audited delete, and other plugin tools to enable immediate action with in the folder tree.

Deliver Value to All Stakeholders

![Delivers The Best Value When All Customer Stakeholder Needs Are Met](image)

Figure 2  Solution has value for many stakeholders
Key functional benefits and features

The primary business value of ClarityNow, as well as its primary functional use case, is grounded in the ability to quickly scan and index massive unstructured data repositories across many platforms. At its core, ClarityNow is a file indexing and reporting engine. It is designed to help an organization define and classify its data repositories based on custom business criteria, rather than simply by file type, date, or ACL. Working in this way, ClarityNow presents a business management view into organizational data repositories through robust classification and reporting tools, while maintaining already in-place security and access restrictions – in other words, visibility without access.

Assign TCO costs and view storage consumption costs

ClarityNow provides a mechanism to assign Total Cost of Ownership values to storage platforms within the enterprise.

Within the GUI main window, it is possible to clearly and simply see average age of data in comparison with costs associated. Managers may assign total cost of ownership values to each volume ClarityNow scans. This provides the business with a way to gauge consumption costs and assist data asset managers to quickly zero in on what is consuming valuable storage.
Management level view of data storage without access

What does visibility without access mean? With ClarityNow, there is no direct user access to the data, only to preset data movement plugins and/or audited delete.

ClarityNow provides the benefit of single pane management view, for multiple network filesystems on most industry standard network storage platforms, using either a web browser client, or a desktop java client which provides greater functionality. Since ClarityNow is building reports based on scan jobs and indexes contained within its own database, ClarityNow has no need to directly access actual data files, leaving all security, AD, and ACL/DACL rules in place.

Out of band file indexing and data movement

The ClarityNow server rapidly builds an index database of the organization by mounting remote shared filesystems and making S3 calls to object data stores. Note that ClarityNow is out of band to the actual data path workflow. All data resources can still be accessed via normal protocols and applications using traditional filesystem mount utilities.

A major key to ClarityNow’s performance is that it stores information on a per folder basis rather than per file. In this way, the index database is populated similarly to the way files on storage tend to be organized in folders or even folder trees.
Cross-platform functionality

ClarityNow indexes at PB scale across multiple platforms. Indexing works on most network filesystem shares as long as the shares can be mounted to a Linux host (ClarityNow server host) using industry standard NFS and SMB protocols, and permissions are set so that the full filesystem can be walked by the ClarityNow server. Read access is all that is necessary. Additionally, ClarityNow can read and index S3 repositories, either local on-premise, or cloud subscription services, using only read access credentials.

ClarityNow also has a plug-in, referred to as a data mover, which can move files or folders from NFS/SMB filesystems to S3 object targets.

Reporting

The main ClarityNow view is an interactive report window for volume and tagged file information. After the initial volume scan, it is possible to view scanned information organized by business context, all customizable through a variety of supported filters and groupings. Report views interactively allow data asset managers to drill down into further detail about how their data assets are organized. Actual file and folder content supporting the volume summary is displayed in the lower pane.

![Chart showing file and folder breakdown](image)

Figure 5  ClarityNow main window reporting on the server UI

There are a series of volume-based reports including:

- Volume aggregate
- All volume summaries
- Tagged files by category
- Error reports and Scan errors
- Files grouped by tagged status
- Files grouped by expiration status
- Volumes by capacity limits status
Depending on tags applied to files and folders, summary reports can be highly customized according to business context. The sample below is a Media & Entertainment example of Volume GB consumed according to various files associated by Movie/Show title. Of importance is the fact that one show title (designated by color bar) has file content scattered across multiple volumes. Tags applied to the ClarityNow index database allow that information to be summarized as show in this sample giving a very easy view of how valuable production storage space is being consumed.

Figure 6   Example of customized reporting using tagged information

Individual Folder/File tracking

For the majority of workflows, ClarityNow stores information on a per folder basis rather than per file. This parallels the way files on storage tend to be organized in folders or even folder trees. Therefore, ClarityNow is able to extract greater efficiency by tracking changes at the folder level for files within the folder.

Some workflows present exceptions to this general rule. For example, multiple streaming video files are sometimes stored together in a folder, despite belonging to different projects or even customers. To address these types of situations, ClarityNow can be configured to also track files and file types individually.
ClarityNow Rich Feature/Functionality

ClarityNow distinguishes itself by combining an impressive set of features and functionality designed specifically to enable successful management and administration of ever growing data assets. A partial list of ClarityNow features includes:

- Easily add new volumes for scanning
- Customizable data discovery by rapid filesystem scan
- Quick summary views of newly added volumes
- Data categorization by tagging
- Extensive autotagging
- Individual folder and file tracking
- Cost basis storage usage reporting
- Emailed alerts
- Plugin enabled file to S3 data transfer
- Cross platform scanning
- PB scale indexing
- LDAP / multi-user enabled
- Scan results available to multiple delegated users
- Handy action cart allowing folder names to be collected for future tag searches

Each functional capability of ClarityNow meets a real-world need. The software was organically developed in response to practical use cases and customer driven pain points.

File indexing via rapid filesystem scan

ClarityNow scans NFS or SMB filesystems which are mounted to the ClarityNow server. Read-only permissions are adequate to allow the system to complete its indexing operation and thereby maintain security. As the server scan walks through the filesystem, it populates a memory resident index database which stores information about the contents of the filesystem, but does not open, read, or store information about the content of the files themselves. ClarityNow gains speed and efficiency by recording information on a per folder basis rather than a per file basis. Exceptions to this rule are possible for certain circumstances.
Storing information per folder permits ClarityNow to keep the records of all folders in memory. This allows for instant access to rolled up information, such as the aggregate size, average age, and last modification date of files in any folder tree. It also allows for fast searches to be performed across all managed storage, without incurring any disk IO.

Filesystem scans may be issued manually by an authorized user, or run as a scheduled job. The scan log indicates how the scan was run and if any errors occurred. For the most part, scan errors relate most often to permission issues within the filesystem.

To deliver greater performance and flexibility, there are two types of scans: Full and Optimized.

- **Full scans**
  
  The first time a storage filesystem is indexed, a full scan needs to be completed. ClarityNow will walk the entire set of filesystems mounted to the server, indexing every folder. This initial baseline scan ensures that everything about the filesystem is known.

- **Optimized scans**
  
  After the full scan completes, subsequent rescans are optimized. ClarityNow does not need to descend into folders where there have been no changes. This significantly increases the speed of indexing filesystems.

By default, ClarityNow stores tracking information about the filesystem on a per-folder basis. Data for individual files is usually aggregated into the parent folder and details are not kept on a per-file basis. Once scanned, all folders and files are tracked within the index database.

Additionally, ClarityNow can track numeric sequences of files within a folder such as uncompressed video image sequences which are common in Media & Entertainment workflows. ClarityNow can be configured to track sequences of files but store them as a single ClarityNow object needing 1KB for the entire sequence, thereby maintaining database and fast search efficiency.

### Auto-tagging

One of the most useful and eye-catching features of ClarityNow is the ability to apply custom tags on folders and, occasionally, specific file assets to help sort data into relevant business context. These tag definitions can be completely arbitrary and based solely on business workflow nomenclature. This capability is useful for correlating related folder/file assets which might be stored on different volumes, across multiple platforms, according to how actual project teams are organized within a business. Summary reports can therefore be built which display total storage consumption by project name, team designation, or even project stage of completion rather than being limited to reports based on simple file extensions eg. How much space is consumed by .tiff files, etc.

Auto-tagging is the system’s way of logically grouping folders and file assets within any given business context and generally occurs whenever ClarityNow scans a filesystem. Auto-tagging rules are built within a configuration file. These rules may be developed on an offline copy of a customer’s filesystem data by using...
an exported debug file. Once the rules are correct and tested, they can be applied to an online index
database.

The ‘autotag.cfg’ file which is in the ‘/usr/local/claritynow/etc’ folder contains the rules by which tags are
applied to folders scanned by ClarityNow indexing. Rules must match on the virtual path which represents
the volume, rather than the physical path which usually is the mount point. Auto-tag rules use regular
expressions to find folders which match, and upon doing so, allow aggregate summary information to be
displayed in the reports window. These reports allow the user to interactively drill down in the chart
summaries to see the contents of everything that matches that tag rule.

Note: Autotag functionality is turned off by default. It is advised that the customer have a discussion with SE’s
or engineering to develop a hierarchical tag strategy before turning this functionality on. An over-prolific and
arbitrary approach to file tagging can have significant impact on the ClarityNow Index database.

It should be emphasized that this capability should be used with prudence and only when the use case makes
it necessary. An over proliferation of arbitrary (non-hierarchical) tags on datasets that exceed 100’s of
millions of files could cause drastic performance issues with the ClarityNow index database – even with using
SSD based volumes for the index database.

Auto-Tag report example

As an example of using auto-tagging to generate a report, the ‘…/etc/autotag.cfg’ file has been modified to
include this sample set of instructions containing a regular expression.

```
set
  match (?i)/testIsilon/projects/(prj\d)_[a-z]+
  max_depth 4
  apply_tag Images/$1
```

This rule will tag items in the ‘projects’ folder within a test network shared filesystem that contains several
sub-project folders. Each project subfolder has been named pr\d1, pr\d2, pr\d3, etc to simulate folder sprawl with
similar subfolder names.
Having run the auto-tag rule from the ClarityNow user interface by clicking on:

```
File → Auto tag → Refresh all Folders
```

A message showing which rules were successful and number of matched items is displayed in the UI main window.

```
Auto tag result

Set 1, rule 1: (?i)testslion/projects/([pr]\d)_[a-z]+  
checked 121 times  
matched 5 times  
time: 0.000 s  

Total rules checked: 121  
Total rules matched: 5  
Total time: 0.002 s (includes enumeration)
```

The successful tag operation is represented by the inclusion of tag icons in the folder tree. These indicate that the indexing database now has entries which allows ClarityNow to track these folder for changes and display rolled up reports.
The ClarityNow user interface main window has an icon which will open the report window as shown. Click the reports icon.

Then click on the ‘Open Report Window’ icon-button that appears.

Here is an example of a report generated from this auto-tag solution. It shows summaries of data storage by project name, even though the project folders are broken out into multiple subfolders. For instance, prj2 summary is comprised of prj2_dev and prj2_docs folders but represented in a single bar.

For this example, the arbitrary name of ‘Images’ was the tag applied in the autotag.cfg rule described above. That tag rule is shown in the left-hand pane beneath the Volumes listings. Clicking on that tag rule will also generate reports derived from the indexing on that rule.
If further detail is required, an administrator can drill down by clicking on the data bar within the graph.

An exhaustive discussion of how to develop Auto-tag rules using regular expressions is beyond the scope of this paper. However, here are some general guidelines and format examples.

There are four main components of building an auto-tag rule. These are denote below and correspond to the structural example shown in the overall format example below.

- **<re>: Regular expression**
  
  This is a Java regular expression (RE) that must match the ENTIRE path (^ and $ are assumed) in order for the tag to be applied.

- **<n>: Search depth**
  
  This is the maximum folder depth allowed for finding a match via the regular expression. Increasing this number will increase the number of directories and sub-directories examined, thus decreasing scan performance. /vol1/folder1 requires a search depth of 2, for example.

- **<cat>: Category**

Figure 13  Customized report based on auto-tag information

Rule based tagging options
This is a tag category which is useful for creating a context around successful matches such as project names, project team names, business units, etc.

- `<tag>`: actual tag name.

This may be a parameter such as $1, $2, $2 etc to be replaced by substrings (indicated by parentheses) from the regular expression used above. $user and $group can also be used to display the folder user name/group in the applied tag.

The overall format of an Auto-tag rule (which needs to be inserted into the autotag.cfg file) follows an indented hierarchy of definitions shown here.

```plaintext
set
  match <re>                     # path to match
    max_depth <n>              # maximum search depth
    [required_tag <cat>/tag>]  # rule will only be used if volume has this tag
  apply_tag <cat>/tag>       # tag to apply
  apply_tag <cat>/tag>
    . . .
  match ...

set
  match ...
```

A set is an arbitrary gathering of matching rules. The configuration file can have multiple sets, and multiple matching rules within a set.

---

**Note: Two important guidelines for ClarityNow rule creation**

1. Match on the Virtual Path
2. Match all the way through to the end of the item, either a directory name or file name

Here are some general expected behaviors when filesystem scan attempts to match rules for each tracked folder or tracked file:

- Sets are checked in order listed
- Rules within a set are checked in order listed
- The first rule within a set to match a tracked item will prevent any following regular expression matches within that set from being examined or applied to the same tracked item.
- After a set has been processed, the next set is processed
ClarityNow relies on a RockDB architecture to store configuration, and tagging information. In order to maintain high performance, ClarityNow requires an SSD tier of persistent storage to house its index database. Variables within the ‘/usr/local/claritynow/etc/clarity.cfg’ file can be used to determine how much of the index is stored on SSD, which allows for older and less used index data to be tiered to standard HDD spinning disk. By default, ClarityNow assumes all index data should reside on SSD.

This structure is key to the fast search capability ClarityNow offers data asset managers. Common time consuming tasks such as searching filesystems for log files, lost license files, or folders accidentally moved to unknown locations, become a simple matter of entering a search string and viewing a near instant file/path listing for any matches found in the database. This completely removes the necessity of going to the storage platform and performing real-time file searches using standard POSIX compliant tools, and avoids the time consumption of searches over heavily populated filesystems.

Additionally, as new file/folder content is added, ClarityNow leverages its optimized filesystem scan process to avoid rescanning already indexed content and only indexes net new files and folders. This is another time saving feature of ClarityNow.
Tactical file and folder movement to archive targets

ClarityNow provides administrators and managers the ability to make manual decisions about where data should be placed based on business context and enact those decisions from within the UI of ClarityNow. It is possible to choose a single file, or an entire folder, for a Copy/Move operation from the folder tree UI once the optional data mover plugin has been installed.

Note that the Transfer menu option shown below will not be visible until the optional data mover plugin is installed and a valid data mover service been configured. At that point, the menu option will appear as shown below. Generally speaking, the same is true for other optional plugins which address other desired functionality; menu items appear in context as plugins are added.

In this case, the ‘Transfer’ target can be either another network shared filesystem, or an S3 target.

Figure 15 Selecting a folder and manually choosing to Transfer data

This topic is covered in more depth in the ‘Anatomy of data move process’ section in this paper.

Delegated user permissions

ClarityNow provides a user/user-group model for fine grained control of a set of user actions. User-groups specify a set of access rights. The user dialog is used to manage users. It is found under

Server -> Access -> Users.
Figure 16  Example of User management dialog

The buttons in the toolbar allow adding, editing, and deleting users. To add a user, click on the ‘add user’ button. To edit a user, select the user from the list and click on the ‘edit user’ button. To delete a user, select the user from the list and click on the ‘delete user’ button.

There are two default users:

The ‘root’ user is the administrator of ClarityNow. This user has no access restrictions of any kind.

The ‘anonymous’ user can be used to provide anonymous access to ClarityNow. By default, the anonymous user belongs to the anonymous group which has read-only access to the server. This account is enabled by default.

The user administrator (or root user) is able to edit and modify permissions for other users with the following options.

<table>
<thead>
<tr>
<th>Option</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can edit general configuration</td>
<td>Allowed to edit configuration items not specifically listed here</td>
</tr>
<tr>
<td>Can edit volume configuration</td>
<td>Allowed to add/edit/delete volume configurations</td>
</tr>
<tr>
<td>Can edit users and groups</td>
<td>Allowed to add/edit/delete users and user groups</td>
</tr>
<tr>
<td>Can request scans</td>
<td>Allowed to request manual scans</td>
</tr>
<tr>
<td>Can edit tag categories</td>
<td>Allowed to add/edit/delete tag categories</td>
</tr>
<tr>
<td>Can create and delete tags</td>
<td>Allowed to add/edit/delete tags</td>
</tr>
<tr>
<td>Can edit folder restrictions and folder tags</td>
<td>Allowed to manually set/remove tags to/from tracked items. Also allowed to set size and time restrictions for tracked items</td>
</tr>
</tbody>
</table>
LDAP integration

ClarityNow LDAP integration can be used to authenticate users against a Microsoft Domain Controller (DC). Authentication against generic LDAP servers (such as ApacheDS) has not been tested and is not supported. LDAP integration supports both LDAP and LDAPS protocols.

The ldap protocol sends cleartext passwords, so use of ldaps is highly recommended.

In order to use the LDAPS protocol, the domain controller must be configured with a certificate. ClarityNow LDAP integration supports self-signed certificates. If a self-signed certificate is used, you will be asked to confirm the use of the certificate when configuring LDAP integration. Any users not configured to use LDAP will be authenticated with their normal ClarityNow Login and password.

Cross platform / Multiplatform support

ClarityNow indexing is capable of cross platform reporting. It is able to work on most network filesystem shares as long as the shares can be mounted to the Linux host serving as the ClarityNow server. Using industry standard NFS and CIFS protocols, volumes are configured for each mounted filesystem. In this way multiple network shares from a variety of platforms may be scanned and indexed on up to PetaByte scale.

In most cases, read permissions are all that is necessary for ClarityNow access. Additionally, ClarityNow can read and index S3 repositories, either local on premise, or cloud subscription services, again as long as
access credentials are provided and read access granted. Access to Azure subscription services is not supported.

Because ClarityNow mounts SMB network shares to the Linux host used for ClarityNow services, there is no support for maintaining ACL/DACL metadata information when transferring files.

**Action cart**

Another functional feature of the ClarityNow user interface is the presence of the Action Cart. The action cart is useful for collecting lists of paths that can have a future action assigned to them as a group.

![Action cart with three files and options presented](image)

Figure 18  Action cart with three files and options presented

Typical uses include putting together a list of paths to attach to an email and collecting a set of paths to flag for archiving or deletion.

**Offline scan log analysis**

A debug file can be used to run offline analytics on the gathered filesystem information. It can be useful for refining autotag rules, researching through report views on tagged information, and cleanup of files ready for archive or delete. The debug file allows the user to import a data set into an offline instance of ClarityNow for review.

After running

Help → Save Debug Info → Full

Copy the saved file to an offline separate instance of ClarityNow for further rule testing and report generation.

1. Unzip the dataset to a location /data/customer_debug
2. Optionally reset the password on the dataset
   a. cd /usr/local/claritynow/scripts
   b. ./server.sh --init --root /data/customer_debug

3. Start the debug instance
   a. cd /usr/local/claritynow/scripts
   b. ./server.sh --root /data/customer_debug

Having taken these steps, ClarityNow will open with the new data set in an offline state.

Note that 3b will run ClarityNow in the foreground. Press ctrl-C to stop the service or run the process in the background.
Architecture

ClarityNow is built upon a traditional client/server model. Therefore, many clients can be running and accessing the ClarityNow server simultaneously. The server is java based and runs a variety of internal services such as volume scanning and indexing. The server is responsible for tagging information about any mounted volumes relevant to any specific business context as determined by the customer.

The ClarityNow user interface UI can be accessed via a simple web view (with limited functionality), or by installing a java based local client (fat client) which provides full server management functionality according to permissions assigned during login.

ClarityNow Server

The server is a software component that is installed on a Linux host. Visit support.dell.com for a list of supported Linux distributions. ClarityNow relies on an ACID compliant, journal-backed, in-memory database for configuration, and tagging information. Keeping this data in-memory yields very high performance, while a change journal allows for reconstructing the database state if a loss of in-memory state occurs due to an unplanned outage.

The ClarityNow server may also fulfill an all-in-one purpose by handling some data mover tasks. However, for performance reasons, it is recommended to move any data mover functionality out to separate linux hosts.

Folder and file data such as size, last modified date, and average age are kept in the index database. Changes to the filesystem structure are recorded occur only during scans, whether manually initiated or scheduled.
Server IndexDB and SSD

ClarityNow server employs a RocksDB, a large key value store, that utilizes a multi-tier persistent storage. It is now a requirement that the indexes for ClarityNow volume scans reside on the ssd tier. There are ten instances of RocksDB within a deployment to allow for higher performance during concurrent scans by multiple clients. A volume is located in only one shard while search and annotation indexes are spread across all DB instances.

Referential integrity is not conserved or maintained at an action by action level. In other words, deletion of a folder within a volume is not immediately represented within the scanned indexes. A weekly maintenance routine removes orphaned index and folder tree entries. The same routine can be run manually as well.

![Figure 19 ClarityNow Server and RocksDB](image-url)
Multiple communication channels to ClarityNow server

There are three main communication channels available with ClarityNow.

- **DRMI** - DataFrameworks enhancements to the Java Remote Method Invocation allows efficient communications between multiple GUI clients and the command line interface tools. The enhancements ensure ClarityNow can easily support firewall security requirements by minimizing the number of needed network ports.
- **JSON** - The JSON interface is used by the Python client API and any direct API access to ClarityNow. This interface provides integration opportunities with external services.
- **HTML** - An optional way to request reports by sending a pre-formatted URL directly from a web browser. ClarityNow server will respond with a bar chart graphical summary report for data corresponding to the URL request.

ClarityNow client and web reports access

The ClarityNow desktop client provides a dashboard view as has been shown earlier in this whitepaper and such as show here again in figure 18. The main window provides an overview of all the mounted and defined volumes as well as some summary pie charts and information about folders and files at the level selected in the folder tree.
The desktop client challenges the user for login credentials to gain access to run reports, scans, auto-tagging, and indexing according to pre-designated permissions assigned by the administrator.

In addition to using the desktop client for reports, it is possible to use a web browser to query the ClarityNow server to gain similar reports. Here is an example of a web browser query on the previously discussed ‘Images’ tag report.

For this example, a direct IP address was used, but a DNS hostname could have been used in similar fashion. Web reports allow access to ClarityNow reports in a browser using a URL. As web reports are available to everyone who can connect to the ClarityNow server with a browser, web reports must be enabled by the ClarityNow administrator in the “server” configuration dialog.
Web report URL format

The basic web report request URL format is as follows:

```
http://<hostname>/reports/<report_type>?<options>
```

Where

- `<hostname>`
  - is the ClarityNow server host name
- `<options>`
  - is a set of options described below.
- `<report_type>`
  - is either 'chart' or 'paths'.

Options are separated by ampersand (&). Some options are in the form of name/value pairs (name=value) while others are flags that are specified without a value.

In the case of conflicting options, the options further to the right will override conflicting options to their left.

Data Mover

ClarityNow offers a data mover plug-in which runs as a separate set of services and enables users to view the index, and then take actions such as data movement. These services can run on the actual ClarityNow server, or on a separate Linux host where they login remotely to the ClarityNow server to receive instructions.

An important part of the ClarityNow design is that its services are considered not in-line from a data path perspective. In all cases, whether indexing, scanning, tagging, or even data movement, regular file and object operations are all still fully accessible from a normal network tools and applications standpoint.

For performance reasons, it may be advisable to build separate Data Mover hosts (nodes) if needing to consistently move large amounts of data from platform to platform.

Examples in this whitepaper use a separately configured data mover host.

Capability of data mover units

The data mover service can be adjusted by modifying the `workers.cfg` file. To fine tune performance, customers can adjust the number of worker threads, although the defaults will work for most lower-volume situations. There can also be several worker nodes handling different segments of the available network shares. ClarityNow server will utilize available data mover nodes to run parallel tasks.

The Data Mover allows files to be transferred:
Between filesystems
- To S3-compatible object storage from a mounted filesystem
- From S3-compatible object storage to a mounted filesystem

Additional data mover notes:

Scanning: ClarityNow can mount anything mounted to the Linux server, but will not be able to scan for Windows ACLs.

Data Movement: Data can be moved from NFS preserving POSIX permissions, but moving content from SMB will not retain ACLs.

Distribution of components on server host and on data mover host

The data mover plugin is comprised of components both on server host and the data mover host which work together to provide the data movement service.

Aside from all the software components installed which make up the ClarityNow server, the ‘cn_data_mover_plugin.cfg’ file directs communication parameters with the data mover service node. The default version of this file allows for largely unsecured communication between server and node. This configuration file can be widely customized to meet security needs.

The data mover node, which is installed via the RPM plug-in package, has fewer software components than the ClarityNow server host. However, the plug-in has its own configuration folder in the '/usr/local/data_mover_workers/' directory. This folder contains the ‘workers.cfg’ configuration file which contains information about S3 buckets, volume information, and directions to connect to the ClarityNow server.

Data Mover service components

The data mover plugin has three main service components which work together to coordinate all data transfer activity. A successful data transfer involves

- ClarityNow server alerts data mover node of transfer task request
- Data mover node verifies source and target paths and space allocations
- Data mover node completes actual data transfer from source to target
- Data mover node updates ClarityNow server indexes with changed file location

The three components involved in transfer activity are:

a) ClarityNow plugin – installed by RPM package on data mover node (or ClarityNow server)
b) Data Mover Service – runs as a system level process
c) Data Mover Workers

- LightWorker - path validation checks, path reservations and disk space checks
- HeavyWorker - Does enumerations, preallocations and file/folder transfer operations

The ‘workers.cfg’ file in ‘/usr/local/data_mover_workers/etc’ has a variety of sections and built in examples. It’s important to take note of two particular sections within this configuration file eg. S3 Bucket information, and Volumes.

Examples from ‘workers.cfg’ file

S3 Bucket information – A section of the ‘/usr/local/data_mover_workers/etc/workers.cfg’ configuration file must define the S3 bucket and have appropriate access credentials, volume name, and URL endpoint for the connection. Below is a screen shot example showing the required fields and samples of key-pair values necessary for accessing ECS S3 storage. Note that it is always recommended to use a load balancer with ECS, so therefore the example directs the data mover node to use a loadbalancer as the “alternate_endpoint”. Values entered into this configuration section should match the values used to initially set up the volumes on the ClarityNow server.

```
S3 buckets:
  "cn-repo":
    "claritynow_volume": "/testECS"
    "access_key": "objectuser"
    "secret_key": "uq0FPATg7shQDDoitZHIyrv7VQOL6DGkCaL5tqU"
    "alternate_endpoint": "https://s3proxy.ecstme.org"
    "use_ssl": False
```

Figure 24  Example of S3 Bucket configuration
Volumes

Volumes must be defined in terms of the volume name which represents the actual mount point and the virtual path name which will be used for autotagging and scan activities. The definition of volumes is also built in the ‘/usr/local/data_mover_workers/etc/workers.cfg’ configuration file. Note: the Levels of indentation in the lines of the config file must be maintained as shown in this example.

```yaml
vol_attributes:
  "Isilon-prod":
    "mount_point": "/ProdVolume"
  "Isilon-dev":
    "mount_point": "/DevVolume"
  "Isilon-SMB":
    "mount_point": "/Marketing"
    "ignore_attributes": True
```

Notice the mounts work with Isilon SmartConnect pools.

Data mover plugin overall capability

The data mover plugin supports

- File to File data movement with a scale out “worker node” architecture
- File to S3 data movement with a scale out “worker node” architecture and preserves “data independence” by leaving a standard S3 object in the destination. This object can be retrieved and consumed independent of any ClarityNow component
- Does not support S3 to S3 data movement, however ECS Sync does support S3 to S3 data movement
Anatomy of data move process

An end-to-end process overview for moving data from a network shared filesystem to an S3 object target would include a number of required steps.

1. ClarityNow server scans and indexes a mounted network shared filesystem
2. The network share is also mounted to a linux host also running the ClarityNow data mover service plugin. (Data Mover Service must be logged into ClarityNow server)
3. A file or folder is manually selected via the ClarityNow user interface
4. ClarityNow server sends instructions to the Data Mover Service
5. After validating file/folder path, Data Mover heavy-worker begins copy job to S3 object target (sending only the data, not ACL/DACL information). No file-stubs are used. The full file is either copied to target or copied with source deleted.
6. Using embedded credentials Data Mover worker logs into cloud subscription service or opens SSL tunnel to recommended load balancer for ECS solution. Data copy/move operation is complete.
Choosing a file or folder

Moving data from one platform to another using the data mover plugin on ClarityNow is not a complicated process once all software components are correctly configured. In its basic sense, ClarityNow allows the users to browse the folder structure, select a file or folder, and choose to transfer it to another platform.

Files and folders may be selected for data transfer (data movement) directly from the main window within the ClarityNow user interface. Right click on any file or folder to see a menu dialog which provides action options for that selection. To immediately begin the transfer process, select ‘Transfer’ from the options provided.

One useful option is to add the selected file or folder to the Actions Cart by selecting ‘Copy to Cart’. In this way an action can be made on several items at the same time.

![Example of choosing a folder for Transfer](image)

Figure 26   Example of choosing a folder for Transfer
Transfer to target

After a Transfer job has been selected to run, a second dialog box is presented which primarily asks which target to send the data. This is the only required field to be addressed. There are other options such as providing a custom destination volume/path. Customers may also choose to require the data mover to transfer files in-sequence to make sure they are written in sequence on the target platform. If the ‘Delete Source’ box is checked, then the Data Copy operation becomes a Data Move (Copy/Delete) operation. Default is Data Copy.

![Transfer dialog box]

Selection of transfer target

Validate data move

For the purposes of this example, and empty ECS Bucket was created. The transfer job illustrated above completed the Data Copy operation to ECS via S3 protocol. Looking at the folder/file tree in the ClarityNow main window, we can see that the volume ‘testECS’ is no longer empty, having 1.46GiB of content now that data transfer is complete. This further validates the Data Move capability of filesystem to S3 object.

![Folder/file tree]

Validate successful data transfer
As described earlier, ClarityNow is fully out-of-band file/object indexing and management. With the file transfer example shown above, we can further validate that the successfully transferred file (or folder) is directly accessible in a regular workflow. Opening an S3 Browser instance, we can see that the file has been transferred to the object storage, and that it is accessible via standard S3 protocol.

![S3 Browser](image.png)

Figure 29  Direct object access using S3 proving ClarityNow is out-of-band
Conclusion

ClarityNow is the go-to resource able to address today’s data asset management needs. Its rich feature set provides immediate value to the business by helping data asset managers and stakeholders to gain a more insightful view of data resources across all of their network storage platforms, and customize that view to be consistent with their unique business context. Optimized scans and fast searches gain back valuable time wasted using traditional methods of locating misplaced files and folders and enable managers to surgically select files for transfer to less costly archive storage platforms such as ECS.