Abstract
A best practices guide containing DataIQ best practices and recommendations for common deployment to help avoid difficulties. Document includes descriptions of some default settings and general performance tips.

March 2020
Revisions

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>December 2018</td>
<td>Initial release</td>
</tr>
<tr>
<td>January 2019</td>
<td>3.0 MR release update</td>
</tr>
<tr>
<td>March 2020</td>
<td>DataIQ V1 Release update</td>
</tr>
</tbody>
</table>

Acknowledgements

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

Author: Lynn Ragan

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.

Copyright © 2020 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. [3/19/2020] [Best Practices Guide] [H17560.2]
# Table of contents

Revisions ........................................................................................................................................... 2
Acknowledgements ................................................................................................................................. 2
Table of contents .................................................................................................................................... 3
Executive summary .................................................................................................................................. 5

1 Volume configuration ............................................................................................................................. 6
   1.1 Volumes: filesystem mount and scanning ......................................................................................... 6
   1.2 NFS Mount – no root squash .............................................................................................................. 6
   1.3 S3 ECS access DataIQ server ............................................................................................................ 7
   1.4 Cloud Providers currently supported ............................................................................................... 7

2 Hostname requirements ............................................................................................................................ 8
   2.1 Matching DNS entry ......................................................................................................................... 8
   2.2 FQDN Hostname during installation .................................................................................................. 8

3 Sizing and Performance .......................................................................................................................... 9
   3.1 SSD for index database ..................................................................................................................... 9
   3.2 Server sizing guidelines .................................................................................................................... 9
   3.3 Physical host machines for best performance .................................................................................. 9

4 Security .................................................................................................................................................. 10
   4.1 Server passwords .............................................................................................................................. 10
   4.2 LDAP and cleartext ........................................................................................................................ 10

5 Configuration .......................................................................................................................................... 11
   5.1 Format of configuration files ........................................................................................................... 11
   5.2 Boolean values and numbers ........................................................................................................... 11
   5.3 Data mover configuration ................................................................................................................. 11

6 Auto-tagging .......................................................................................................................................... 12
   6.1 Pre-existing customer tags .............................................................................................................. 12
   6.2 Regex rules built on virtual path ...................................................................................................... 12
   6.3 Match regex rule to end of search item ........................................................................................... 12
   6.4 Auto-tag rule changes and rule hierarchy ..................................................................................... 12

7 File/Folder user configurable threshold restrictions ............................................................................ 13
   7.1 Maximum objects in a folder ............................................................................................................ 13

8 General System and Network ................................................................................................................. 14
   8.1 Avoid placing DataIQ server on multi-purpose servers ................................................................. 14
   8.2 Separate dedicated DataIQ server and DataMover ....................................................................... 14
   8.3 Correct time ................................................................................................................................... 14
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.4 S3 default ‘Offline View’</td>
<td>14</td>
</tr>
<tr>
<td>9 Performance</td>
<td>15</td>
</tr>
<tr>
<td>9.1 Scan threads</td>
<td>15</td>
</tr>
<tr>
<td>9.2 Indexing and search</td>
<td>15</td>
</tr>
<tr>
<td>9.3 Bare metal</td>
<td>15</td>
</tr>
<tr>
<td>A Appendix: Plugins</td>
<td>16</td>
</tr>
<tr>
<td>A.1 DataMover Plugin</td>
<td>16</td>
</tr>
<tr>
<td>B Technical support and resources</td>
<td>19</td>
</tr>
</tbody>
</table>
Executive summary

DataIQ offers critical insight for data managers tasked with organizing ever growing repositories of unstructured data. As a management tool, DataIQ has many configuration options and capabilities which allow it to provide a single pane of glass view across multiple storage platforms. As such, it is designed to be flexible so that DataIQ can be deployed to address a variety of workflows and provide benefit to the customer within their specific business context.

Some commonly encountered best practices have been developed based on practical application of the DataIQ tool to real life customer circumstances. Those recommendations are collected in this guide for reference and to provide assistance to presales effort to introduce DataIQ into new customer environments.
Volume configuration

1 Volume configuration

1.1 Volumes: filesystem mount and scanning

DataIQ scans and indexes network attached filesystems including NFS and SMB. Read-only privileges can be used to preserve security when mounting network shared filesystems to the DataIQ server. Filesystems are mounted manually, using standard Linux mount commands, at the commandline on the DataIQ server host and configured as volumes using the DataIQ WEBUI. NOTE: Volume naming is case sensitive in all situations and must match exactly for both DataIQ server WEBUI, commandline configuration files and data mover plugin.

Because DataIQ uses a containerized microservices architecture all mount points used to create volumes for scanning must be mounted under ‘/mnt’ along side the mount point for the SSD filesystem used to house the Index database. For example:

/mnt/ssl
/mnt/NFS-test1
/mnt/NFS-Prod1
/mnt/SMB-Prod2

For external datamover hosts, it is best if the mountpoints match those configured on the DataIQ server. Mounted network shares must be noted in the ‘workers.cfg’ file which can be located under ‘/usr/local/data_mover_workers/etc’. It is important the volume names must match Volumes are configured by editing the ‘data_mover_service.cfg’ file. Config file should be formatted similar to this example – precise indenting and spacing is critical:

The config file has a vol_attributes: section. The next line below provides the Volume name in quotes which should match the Volume name on the DataIQ server. The third line uses the key-value name "mount_point": which must be present and identify the full path mountpoint value.

```
vol_attributes:  
  "NFS-prod":  
    "mount_point": "/nfsProd1"
  "NFS-test1":  
    "mount_point": "/DevVolume"
  "Isilon-SMB":  
    "mount_point": "/SMB-Prod2"
    "ignore_attributes": True
```

1.2 NFS Mount – no root squash

In order for DataIQ to have NFS permissions access required for Scanning and Indexing, it is important to properly format the NFS exports. In OneFS, under the Protocols tab, there is a place to specify:

- NFS Clients
- Always read/write clients
- Root Clients
At the very minimum, both NFS Clients and Root Clients should have an entry for the IP address of the DataIQ server. For full access, so that plugin functionality also works, ‘Always read/write clients’ should also include an entry for the IP address of the DataIQ server.

Previous versions of the software suggested that ‘no root squash’ should be used in their mounts and/or exports. ‘root’ user would encounter permission errors when trying to scan if this flag is not set when NFS filesystems are mounted. However, Isilon best practices identified this setting as a potential security risk and deprecated the practice. For isolated test labs, in a trusted environment, this may still be a quicker option for test purposes.

1.3 S3 ECS access DataIQ server
Following installation, it is necessary to set up a credentials file which DataIQ server will use to access the ECS S3 endpoint URL.

Edit /opt/dataiq/maunakea/aws/credentials file and place the following inside the file and save it:

```bash
[ecs]
aws_access_key_id=<your bucket’s access key>
aws_secret_access_key=<your bucket’s secret access key>
```

The [ecs] identifier can be any value. This is what is used in the webUI when configuring ECS targets.

The two entries correspond to ‘Username’ and ‘Secret Key’ for any S3 user on ECS.

The [default] shown above can be any keyword value (such as ‘corp’, ‘ECS target’, etc), however, it must match the value referenced in the S3 endpoint configuration UI window.

1.4 Cloud Providers currently supported

DataIQ supports cloud provider using S3 protocol. Primarily, S3 is suitable only for ECS on-premise, ECS cloud hosted, and AWS. In addition, GCP (Google Cloud) is also supported.
2 Hostname requirements

It is required that the hostname be set at the OS level prior to beginning installation of the DataIQ software. Hostname must be set in Fully Qualified Domain Name format (FQDN) such as dataiq.mycompany.com.

2.1 Matching DNS entry

Ideally, a matching entry should also be created as an A-level record in the DNS domain for the customer network.

2.2 FQDN Hostname during installation

During installation of the DataIQ software, the FQDN and the IP address of the DataIQ server must be entered.
3 Sizing and Performance

Sizing a DataIQ solution must always take into account the complexity of the customer storage environment as well as the total number of files/folders and objects per volume.

**Note:** Mass-filesystem migrations is not an intended use case for DataIQ, high volume performance benchmarks have not been established.

Some general architecture concepts may be useful to help size a customer solution. For instance, DataIQ may have several separate data mover hosts all operating together to facilitate a single storage platform thus enabling a scalable model. Workloads can be distributed across the external data mover nodes in a scale-out fashion.

3.1 SSD for index database

DataIQ Server requires an SSD tier which must be mounted under /mnt/ssd. This partition or mounted volume of SSD should be sized per the following guidelines:

250GB of SSD for each 1 Billion objects scanned.

**Note:** Customers should be mindful that extensive tagging also adds to the index database and increases the rate at which this resource is consumed. It can also have the effect of slowing down database functions.

3.2 Server sizing guidelines

Sizing a DataIQ server RAM no longer depends on file and directory count since the index folder has been moved to SSD. As noted above, SSD is now a requirement for the index database.

Below are some DataIQ server sizing guidelines:

- DataIQ index folder must be on SSD mounted volume
- Minimum RAM is 64 GB / additional Memory may be needed if customer predicts large number of concurrent users/scans/searches
- 8 CPU cores for virtualized environments

The object database consists of 1K per directory. Tag information is contained within that 1K for each folder.

3.3 Physical host machines for best performance

Both indexing server and data mover hosts can be run on VM’s in many situations. However, for production, the data movers are recommended to be physical machines to enable better performance (dedicated networks and CPU resources) for high object count or high user concurrency environments. Additionally, physical hosts may be advisable for data mover specific hosts configured for network pipes and using high CPU for checksum calculations.

In most business cases, Virtualized deployments using the above resource guidelines will be sufficient to get the customer up and running.
4 Security

4.1 Server passwords
Customers are advised to change the root password of DataIQ to improve the security of the system. Server passwords are separate from the native server host passwords.

Also, anonymous user/groups should be disabled, unless a policy exists to allow anonymous user login.

4.2 LDAP and cleartext
DataIQ provides for AD/LDAP authentication and inherited AD groups. Using LDAP sends the administrator password in clear text. LDAPs does encrypt passwords, however AD trusted CA cert must be imported first into the DataIQ server using an onboard script.

Use of LADPS is highly recommended.
5 Configuration
Both the DataIQ server and data mover plugins are configurable in most cases via the DataIQ webUI.

5.1 Format of configuration files
When editing DataIQ datamover plugin YAML configuration files, ensure you are only using spaces and not tabs. Also ensure the indentation matches exactly. Config file should be formatted similar to this example:

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

5.2 Boolean values and numbers
Keep Boolean values and numbers unquoted, but all other values should be in quotes.

5.3 Data mover configuration
Volume names configured on the data mover host must match the DataIQ server volume names exactly. Also, the volume name must begin with a ‘/’.
6  **Auto-tagging**
Auto-tagging is the system’s way of tracking items and generally occurs whenever DataIQ 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.

6.1  **Pre-existing customer tags**
Custom tags do not flow from filesystem to ECS. DataIQ does not absorb existing tags which may be assigned to data resident on other systems. The DataIQ auto-tag engine is responsible for assigning tags and tag values.

6.2  **Regex rules built on virtual path**
Always build rules based on the Virtual path for the volume rather than the actual mount point for the volume.

6.3  **Match regex rule to end of search item**
Match the regex rule all the way through to the end of the item, either a directory name or file name.

6.4  **Auto-tag rule changes and rule hierarchy**
Sometimes tags created by Auto-Tagging are lost due to an alteration of the of the matching rules applied in the Auto-Tag configuration section of the DataIQ webUI. In such cases, some auto-tags that had been placed on items may disappear. Tags will only remain on an item if the auto-tag rules have not changed when the next scan occurs. Also, if the order of rules/patterns in the autotag.cfg file changes, then some tags may be lost. This is because a path will only be tagged with the first regex pattern and rule that it matches even though multiple rules may apply to the same file. All patterns/rules that follow in the same rule set will be ignored after a pattern/rule has already been applied from within that set.
7 File/Folder user configurable threshold restrictions

7.1 Maximum objects in a folder

The maxObjectsInFolderEnumeration option is a guardrail for the maximum number of objects in a single folder. Most filesystems have a limit to the number of objects that are allowed in a single folder. When this number is exceeded, a number of issues may arise. Most notably is that it may become virtually impossible to list the contents of a folder. Under normal indexing operation, DataIQ will abort folder enumeration when it encounters more than one million objects in a single folder. The thread that is performing that enumeration will throw an error and move on to the next item to be indexed. This is done to indicate to the administrators that there may be an underlying issue with the filesystem. Performance often suffers greatly when enumerating folders that have too many files. This setting can be adjusted to ensure the enumeration will continue. An administrator should expect significant scan performance issues when there exists an unbounded number of objects in a single folder. When this occurs, the administrator should fix the underlying problem by consulting their filesystem vendor. This option defaults to one million.

\[ \text{maxObjectsInFolderEnumeration} = 1000000 \]

It should be noted that this section is tunable for differing work circumstances in which a separate patternStartLimitMode variable is provided to customize the pattern starting points within the filenames that are being counted.
8 General System and Network

DataIQ operates on standard TCP ports, most of which are configurable by editing the configuration files. It's recommended to leave the ports to their default assignments. These include:

- DataIQ WEBUI (HTTP port) = 443
- Data mover plugin communication to DataIQ Server = 443

8.1 Avoid placing DataIQ server on multi-purpose servers

On a multi-purposed server, hosting DataIQ as well as other applications may experience issues caused by conflicts on ports used by DataIQ, and some DataIQ plugins (like the Data Mover). It is recommended to use a dedicated server host for each instance of DataIQ (server or data mover host) in order to avoid potential network port or resource conflicts.

8.2 Separate dedicated DataIQ server and DataMover

Separate dedicated DataIQ server and external datamover worker node(s) are considered a best practice. The DataIQ server is essentially a high-performance database engine. Its dedicated functions are Scanning, Indexing, Tagging, Search, WebUI, Authentication and assigning datamove jobs. Datamover jobs are, by definition, IO intensive. While DataIQ server does have a single DataMover worker thread (when DataMover is installed), it is used for setup and testing primarily. For customers who will very rarely use DataMover, this may be sufficient. However, if the workflow use case is for more regular and active Data Move activity, then it is highly recommended to have an external DataMover worker node.

8.3 Correct time

Especially for the Data Mover, and for DataIQ itself, when dealing with S3 data stores that use SSL, make sure that their servers have the correct date, time and time zone. SSL certificate authenticated encryption is very time sensitive.

It is recommended that both DataIQ and ECS S3 targets be synchronized to the same NTP sources and rely on NTP (Network Time Protocol) services.

8.4 S3 default ‘Offline View’

Whenever a S3 volume is selected, the “offline view” message will appear by default. The “offline view” will disappear when another filesystem based volume had been selected. This is confirmed default behavior. The intent is to not index S3 volumes as frequently (especially in AWS) as the indexing process can introduce very nominal AWS charges.
9 Performance

Some performance factors are tunable in DataIQ.

9.1 Scan threads

Performance is a configurable factor for both the DataIQ server and data mover plugin (dedicated data mover host) in terms of worker threads which may be optimized. These may be adjusted by altering the configuration files, or temporarily on a job by job basis through the UI. The default is a worker thread setting of ‘1’. It is recommended to change this to 10 in order to see some general performance gains. This setting should be safe for most situations. This setting can be configured using the webUI by editing the Volume configuration. In this way, the threads settings can be adjusted on a per-volume basis. The thread value can be applied to either or both Scheduled Scan Threads and Manual Scan Threads.

Scheduled Scan Threads = 10 (default is a value of 1)

Manual Scan Threads = 1

9.2 Indexing and search

During a volume scan, the indexing database records information pertaining to the file structure, gathering data about file and folder names and paths. It is optimized in such a way as to allow for searches on the first few characters of either files or folder names. This reduces significantly the size of the index database.

If the customer wishes to search on front-end wildcards such as ‘*.iso’ or ‘*team.doc’ then the scanning must be adjusted before the volume is added to DataIQ. The ‘patternStartLimitMode’ controls this behavior and the current default value is ‘patternStartLimitMode=3’. This variable can be adjusted back to the prior setting, which allows search on front-end wildcards, by placing the following line in DataIQ Settings / Data Management Configuration / Advanced Options:

patternStartLimitMode=2

Note that this will slow scans and searches significantly on large-count customer systems. It will also swell the RocksDB (scan+index DB) by many times. Instead of one search index entry per file and folder, there will potentially a dozen or more.

Also note that changes do not clean the prior index work, so it is best to set this upon install and before scanning. If done after scanning, each volume can be deleted and re-added. However, this will force the DB to clean up, and will be slower on giant installs than just removing the SSD DBs and restarting DataIQ.

It is highly recommended that customers request a technical discussion with their account team SE and possibly support to weigh the benefits/risks of adjusting this setting.

9.3 Bare metal

In order to achieve the most consistent and optimal performance, DataIQ is considered a bare metal solution for most highly concurrent use cases. DataIQ does function well in a virtualized environment but is limited in terms of scalability due to the reality of shared resources rather than dedicated resources.
Appendix: Plugins

A. DataMover Plugin

DataMover may be installed on the DataIQ server for test purposes. However, best practice is to install DataMover on a separate VM/Host worker node in order to isolate data IO traffic from the CPU-intensive database work the DataIQ server performs. It is still necessary to install the data mover plugin on the DataIQ server first so that the service is running and available for external worker node call-in.

To install the DataMover plugin/service, the datamover tarball must be extracted and installed on the DataIQ server first.

1. Upload the dataiq-datamover-installer-xxx.tar.gz to /root on DataIQ server.
2. `tar -zxvf dataiq-datamover-installer-xxx.tar.gz`
3. `cd /root/dataiq-datamover`
4. `./install_plugin.sh`

Instructions for installing DataMover plugin on a separate external worker node/host. The RPM bundle is only for the external worker nodes intended to run the datamover plugin.

1. Install standard CentOS 7 Linux distribution (CentOS 8 is not supported at this time)
2. Pull down the 'data_mover_workers-1.0-46.x86_64.rpm'
3. Perform the following

Upload the data_mover_workers-1.0-46.x86_64.rpm to /root/ on the DataMover worker node
Run `rpm -Uh data_mover_workers-1.0-46.x86_64.rpm`

Depending on your CentOS build you may need to add python setup tools and a few other packages.

```
yum install python-setuptools
yum install epel-release
```

OR

```
yum install python-pip
pip --version
pip install --upgrade pip
pip install python-setuptools
pip install python-dateutil
```

Then try installing the datamover workers rpm:
rpm -Uh data_mover_workers-1.0-46.x86_64.rpm

External DataMover workers require configuration via editing config text files. Care must be taken to maintain the YAML format that is in the provided in the config file: workers.cfg. The config file is located:

Cd /usr/local/data_mover_workers

Configuration file is located in the /usr/local/data_mover_workers/etc folder

Cd /etc

Edit `workers.cfg`

Uncomment lines beginning with

- "data_mover_host"
- "data_mover_port"
- "data_mover_password"
- "vol_attributes" (and indented lines below it)
- "S3 buckets" and indented lines below it (if using S3 buckets at all).

Note:

Data mover host needs to have IP of the DataIQ instance.

Data mover port needs to be 443.

Data mover password needs to be the password that appears in the config in Edit configuration for the Data Mover plugin.

Vol_attributes needs to have the local mount points for the volumes used for transfers.

The allow_preallocation and ignore_attributes and min_remaining_space values can be overridden by submitting updates to config via UI. The mount_point will not be overridden.

For mounting multiple volumes it is required to have only one `vol_attributes` heading. Ex.

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

For SMB/CIFS mount points, it likely that the "ignore_attributes": True Boolean needs to be uncommented in order for correct DataMovement operation.

Path to configuration file: /usr/local/data_mover_workers/etc/workers.cfg

Note: Be sure to change the Data Mover password to the same password that the Data Mover service is using. Password can be modified using the DataMover plugin edit option on the DataIQ server as shown:
Under normal circumstances, Scroll down to the Password section and replace ‘CustomizeMe’ with your password which matches the password given to the external DataMover worker node. However, for the purposes of this lab it is important to LEAVE THE PASSWORD AS CustomizeMe. The security module has not been fully developed and is in process. Upcoming versions of the DataMover (soon to be released) will provide for password customization.

It is also important that the Volumes definitions on the external Datamover workernode be exactly copied into the Volumes definitions on the DataIQ server. Syntax is important.

**Note:** Any time you alter the configuration file, you will need to run:

```
systemctl restart data_mover_workers
```

To see the status of the data_mover_workers service, run:

```
systemctl status data_mover_workers
```

To stop the data_mover_workers service, run:

```
systemctl stop data_mover_workers
```

To start the data_mover_workers service after it has been stopped, run:

```
systemctl start data_mover_workers
```

Mount your filesystems (NFS, SMB) on both the DataIQ server host and on the DataMover workernode giving them identical Volume names and identical Mount points.

```
hop-isi-z.solarch.lab.emc.com:/ISOs 161T 132T 28T 83% /mnt/NFSisz
//10.246.156.183/Data 597G 377G 221G 64% /mnt/testCIFS
hop-isi-x.solarch.lab.emc.com:/ifs 108T 20T 85T 20% /mnt/NFSisix
```
B  Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Storage technical documents and videos provide expertise that helps to ensure customer success on Dell EMC storage platforms.