

Project2

Contents

Chapter 1. Cover Page.....	1
Chapter 2. About this publication.....	2
Contacting IBM StoredIQ customer support.....	2
Chapter 3. IBM StoredIQ components.....	3
Solution components.....	3
Gateway.....	3
Data Servers.....	3
Application stack.....	4
Elasticsearch cluster.....	4
Applications of IBM StoredIQ.....	4
IBM StoredIQ Data Server.....	4
IBM StoredIQ Administrator.....	5
IBM StoredIQ Data Workbench.....	8
IBM StoredIQ Insights.....	10
IBM StoredIQ Policy Manager.....	10
IBM StoredIQ Desktop Data Collector.....	11
Chapter 4. Planning for deployment.....	12
Planning for Deployment.....	12
Open Virtual Appliance (OVA) configuration requirements.....	13
Network and port requirements.....	15
Open ports for desktop client access to the data server.....	18
Environment sizing guidelines.....	19
Stack-provisioning prerequisites.....	20
License usage metrics.....	21
Security.....	22
Chapter 5. Deploying IBM StoredIQ.....	25
Deploying IBM StoredIQ.....	25
Deploying the virtual appliances.....	25
Configuration.....	26
Deploying IBM StoredIQ on Microsoft Hyper-V.....	26

sample procedure.....	27
Chapter 6. Information.....	29
Chapter 7. Glossary.....	30
Glossary Terms.....	30
Index.....	a

Chapter 1. Cover Page

The following is the hierarchy used for this project

- Project2
 - resources
 - images
 - pdfs
 - topics
 - concepts
 - references
 - tasks
 - glossary
- There are two maps: the main map is Project2.ditamap
- a secondary map called infomap.ditamap contains keys for conkeyref
- a reference file called source.dita contains the conkeyref references
- a reference file called restofdocument.dita is the destination file for internal links where links occurred in the source document to pages we haven't transcribed.
- Images and notes were inserted using conkeyref.
- another secondary map called indexmap.ditamap was created for Index Terms, linked to Project2.ditamap
 - conkeyref was used in the prolog in autoclassification.dita and licenceusagemetrics.dita to link to the index.
 - product metadata was added to the Project2.ditamap
 - copyright metadata was added to sources.dita

Chapter 2. About this publication

About this publication

IBM StoredIQ Deployment and Configuration Guide provides information about how to plan, deploy, and configure the IBM StoredIQ product.

Contacting IBM StoredIQ customer support

For IBM StoredIQ technical support or to learn about available service options, contact IBM StoredIQ customer support at this phone number:

- 1-866-227-2068

Or, see the Contact IBM web site at <http://www.ibm.com/contact/us/>.

IBM Knowledge Center

The IBM StoredIQ documentation is available in [IBM Knowledge Centre](#)

Contacting IBM

For general inquiries, call 800-IBM-4YOU (800-426-4968). To contact IBM customer service in the United States or Canada, call 1-800-IBM-SERV (1-800-426-7378).

For more information about how to contact IBM, including TTY service, see the Contact IBM website at <http://www.ibm.com/contact/us/>.

Chapter 3. IBM StoredIQ components

IBM StoredIQ components

The IBM StoredIQ solution consists of these components: the application stack, the gateway, the data server, and optionally the Elasticsearch cluster.

Solution components

IBM StoredIQ provides three solution components: the gateway, data servers, and application stack (AppStack)

Gateway

The gateway communicates between the data servers and the application stack. The application stack polls the gateway for information about the data on the data servers. The data servers push the information to the gateway.

Data Servers

A data server obtains the data from supported data sources and indexes it. By indexing this data, you gain information about unstructured data such as file size, file data types, file owners.

The data server pushes the information about volumes and indexes to the gateway so it can be communicated to the application stack. Multiple data servers feed into a single gateway.

Data servers can be categorized in two types: DataServer - Classic and DataServer - Distributed. A data server of the type DataServer - Classic uses the embedded PostgreSQL database for storing the index. With a data server of the type DataServer - Distributed, the index is stored in an Elasticsearch cluster. Data servers of this type also provide better performance in search queries. They can manage much larger amounts of data than data servers of the type DataServer - Classic, thus making the IBM StoredIQ deployments more scalable.

You can have both types of data servers in your IBM StoredIQ deployment.

In addition to completing standard administrative tasks, administrators can deploy the IBM StoredIQ Desktop Data Collector and index desktops from the data server.

Application stack

The application stack provides the user interface for the IBM StoredIQ Administrator, IBM StoredIQ Data Workbench, IBM StoredIQ Insights, and the IBM StoredIQ Policy Manager products.

The synchronization feature for integration with a governance catalog is also part of the application stack.

Elasticsearch cluster

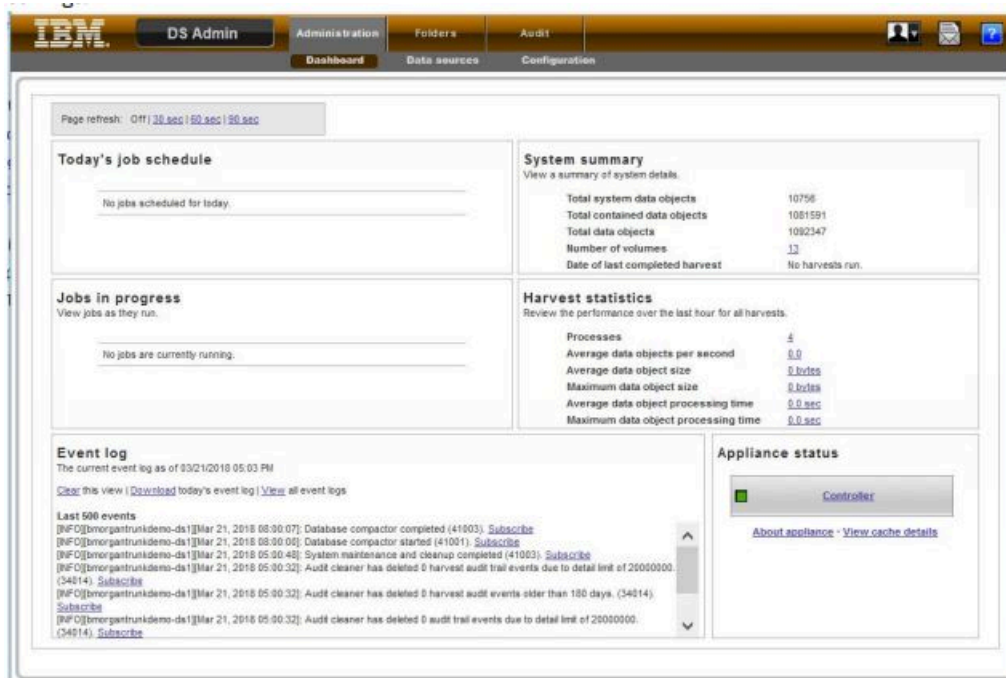
The Elasticsearch cluster attached to a data server of the type DataServer - Distributed provides a single data store for all metadata and content of harvested objects. Indexed data is distributed automatically across the nodes in the cluster. Indexing and queries are load-balanced across all nodes. Nodes can be added dynamically without downtime and the indexing process can use these newly added nodes without further setup.

Applications of IBM StoredIQ

IBM StoredIQ provides interface applications that help fulfill its solution goals.

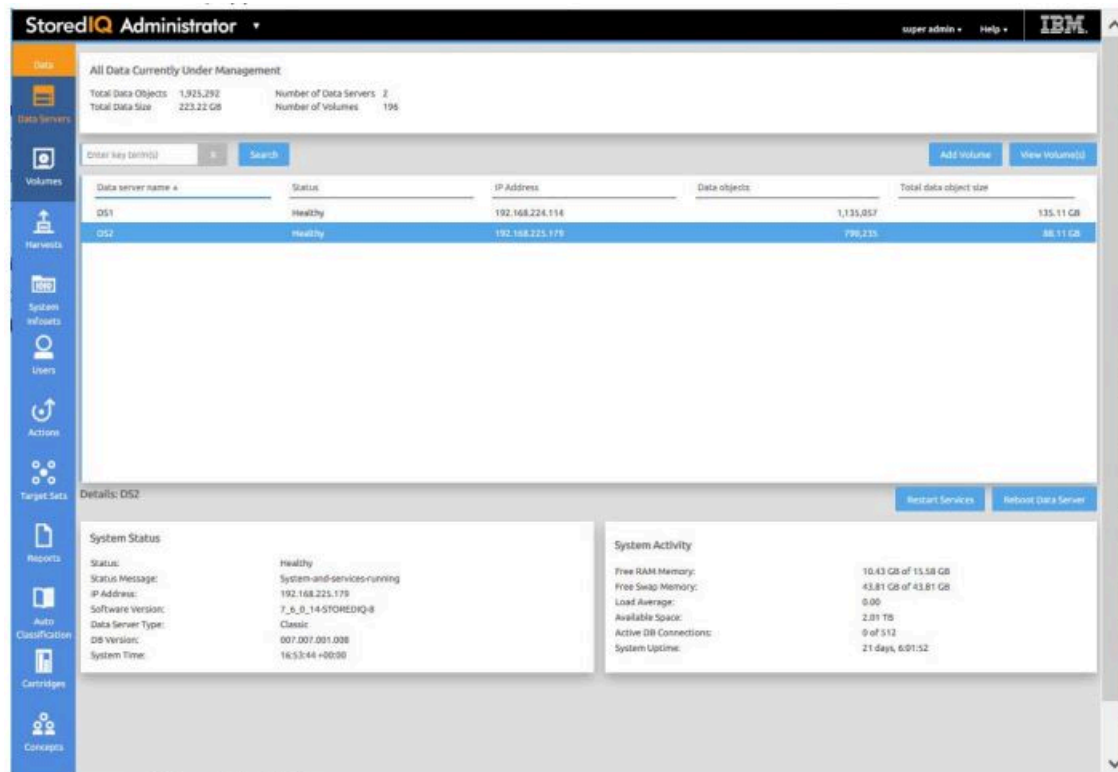
IBM StoredIQ Data Server

IBM StoredIQ Data Server user interface provides access to data server functionality. It allows administrators to view the dashboard and see the status of the jobs and system details. Administrators can manage information about servers and conduct various configurations on the system and application settings.



IBM StoredIQ Administrator

IBM StoredIQ Administrator helps you manage global assets common to the distributed infrastructure behind IBM StoredIQ applications.



IBM StoredIQ Administrator provides at-a-glance understanding of the different issues that can crop up in the IBM StoredIQ environment. These views are unique to the IBM StoredIQ Administrator application as they provide an overview of how the system is running. They allow access to various pieces of information that are being shared across applications or allow for the management of resources in a centralized manner.

The administrator is the person responsible for managing the IBM StoredIQ. This individual has strong understanding of data sources, indexes, data servers, jobs, info sets, and actions. This list provides an overview as to how IBM StoredIQ Administrator works:

Viewing data servers and volumes:

Using IBM StoredIQ Administrator, the Administrator can identify what data servers are deployed, their location, what data is being managed, and the status of each data server in the system. Volume management is a central component of IBM StoredIQ. IBM StoredIQ Administrator also allows the Administrator to see what volumes are currently under management, which data server is responsible for that volume, the state of the volume after indexing, and the amount and size of information that is contained by each volume. Administrators can also add volumes to and delete volumes from data servers through this interface.

If IBM StoredIQ is configured for integration with Information Governance Catalog, the Administrator can also manage which volumes are published to the governance catalog.

Scheduling harvests:

Harvesting, which can also be referred to as indexing, is the process or task by which IBM StoredIQ examines and classifies data in your network. Using IBM StoredIQ Administrator, harvests can be scheduled, edited, and deleted.

Creating system infosets:

System infosets that use only specific indexed volumes can be created and managed within IBM StoredIQ Administrator. Although infosets are a core component of IBM StoredIQ Data Workbench, system infosets are created as a shortcut for users in IBM StoredIQ Administrator.

Managing users:

The user management area allows administrators to create users and manage users' access to the various IBM StoredIQ applications.

Configuring and managing actions:

An action is any process that is taken upon the data that is represented by the indexes. Actions are run by data servers on indexed data objects. Any errors or warnings that are generated as a result of an action are recorded as exceptions in IBM StoredIQ Data Workbench.

 **Note:** Actions can be created within IBM StoredIQ Administrator and then made available to other IBM StoredIQ applications such as IBM StoredIQ Data Workbench.

Managing target sets:

Provides an interface that allows the user to set the wanted targets for specific actions that require a destination volume for their actions.

Reports:

IBM StoredIQ Administrator provides a number of built-in reports, such as summaries of data objects in the system, storage use, and the number of identical documents in the system. You can create custom reports, including Query Analysis Reports for e-discovery purposes, and automatically email report notifications to administrators and other interested parties.

Autoclassification

Automated document categorization, what IBM StoredIQ refers to as autoclassification models, integrates the IBM® Content Classification's classification model into the IBM StoredIQ infoset-generation process. Data Experts can use IBM Content Classification to train a classification model, which is then registered with IBM StoredIQ Administrator. The registered classification model can

be applied to an existing infoset in IBM StoredIQ Data Workbench to generate new metadata for the objects in the infoset. Metadata can be used in rule-based filters to create new infosets.

Cartridges

Cartridges are compressed files that contain analysis logic. When you add a cartridge to IBM StoredIQ AppStack, it can detect new data in documents during indexing and make these new insights searchable. For example, a sensitive pattern cartridge can enable IBM StoredIQ to detect passport numbers, phone numbers, and other IDs.

To apply the analysis logic contained in the cartridge, you must run a Step-up Analytics action that uses the cartridge on an infoset. IBM StoredIQ examines all documents in the infoset, applies the analytics, and then stores the analysis results in the IBM StoredIQ index.

Managing concepts:

Provides the ability to relate business concepts to indexed data.

Managing Mule scripts

Helps you to create Mule scripts and upload script packages. These Mule scripts are used by IBM StoredIQ Policy Manager to create policies using the automation workflow.

DataServer - Classic:

Data servers can be categorized in two types: DataServer - Classic and DataServer - Distributed. DataServer - Classic refers to the regular data servers. It uses either the current PostgreSQL or Lucene index as an index.

DataServer - Distributed:

The distributed data server uses an Elasticsearch cluster instead of an embedded Postgres database. It increases the scalability and flexibility of the IBM StoredIQ deployment in a way that it can manage much larger amounts of data. Without adding more data servers, data that is managed by the IBM StoredIQ deployment can be increased by adding new nodes to the Elasticsearch cluster. Search queries perform better on DataServer - Distributed.

Connector API SDK:

A connector is a software component of IBM StoredIQ that is used to connect to a data source such as a network file system and access its data. Using IBM StoredIQ Connector API SDK, developers of other companies can develop connectors to new data sources outside the IBM StoredIQ development environment. These connectors can be integrated with a live IBM StoredIQ application to index, search, manage, and analyze data on the data source.

IBM StoredIQ Data Workbench

Big data is a pervasive problem, not a one-time occurrence. It is easy for most companies to realize that big data is problematic, but it is hard to identify what problems they have. Big data is all about the unknown, but the unknown cannot be off limits. IBM StoredIQ Data Workbench can help you learn about your data, make educated decisions with your most valuable asset, and turn your company's most dangerous risk into its most valuable asset.

Name	Total objects	Infoset size	Composition	Created	Type	Description
All Data Objects	1,925,292	223.22 GB	Mixed Level		System	All data objects.
All objects from SP (2016...	1,781	242.63 MB	Mixed Level	2015-12-13 11:44 AM	User	
All System-Level Objects	447,393	115.69 GB	Top Level		System	All system-level objects.
big12 ch2	423	37.92 MB	Mixed Level	2016-03-29 9:25 AM	System	
big12 ch2 user	423	37.92 MB	Mixed Level	2016-03-29 9:51 AM	User	
bmorgan-a ds1	4,273	5.22 GB	Mixed Level	2016-03-21 8:36 AM	System	
bmorgan-e ocr	57	160.11 MB	Top Level	2017-02-02 2:15 PM	System	
box2logesh	397	275.75 MB	Mixed Level	2016-03-22 11:40 AM	User	
bug 9168	17	157.96 MB	Top Level	2017-02-02 2:21 PM	User	
Collaborator Role Contains ...	46	14.52 MB	Top Level	2015-12-13 2:38 PM	User	
Collapsed - All objects from...	915	179.03 MB	Top Level	2015-12-13 11:47 AM	User	
DS1 = collaborator login na...	76	15.85 MB	Top Level	2015-12-13 2:28 PM	User	
DS1 all objects PS nimmo8	58	3.28 MB	Mixed Level	2015-12-13 11:05 PM	User	

IBM StoredIQ Data Workbench is a data visualization and management tool that helps you to actively manage your company's data. It helps you to determine how much data you have, where it is, who owns it, and when it was last used. When you have a clear understanding of your company's data landscape, IBM StoredIQ Data Workbench helps you take control of data. You can make informed decisions about your data and act on that knowledge by copying, copying to retention, or conducting a discovery export.

Here are just some examples of how you can use IBM StoredIQ Data Workbench.

- You need to find all company email that is sent from or received by Eileen Sideways (esideways@thecompany.com). You can use IBM StoredIQ Data Workbench to find all email and then copy that data to a predefined repository. You can also use IBM StoredIQ Data Workbench to find all of the esideways@thecompany.com email that occurred between specific dates and then make that email available for review.
- As an administrator, you want to rid your networks and storage of unused data. You can use IBM StoredIQ Data Workbench to find all files that were not modified in more than five years.
- You want to find all image files that are created in 2007. Not only can IBM StoredIQ Data Workbench find all image files that were created in 2007. It also shows how much space they occupy on your network.

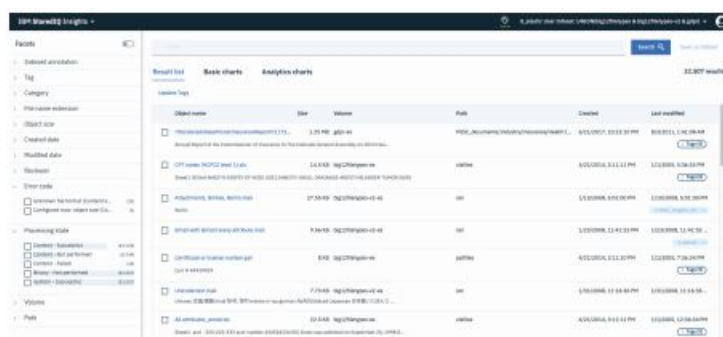
- A user needs to understand how data about Windows is being retained. Using IBM StoredIQ Data Workbench, you can provide that user with a visual overview of the number of objects that are retained and a breakdown of files per data source. Additionally, you can apply overlays to show the user if those files contain forbidden information such as credit-card numbers or Social Security numbers.
- If IBM StoredIQ is configured accordingly, you can select the info sets and filters that are published to the governance catalog for unified governance of structured and unstructured information. When integrating with Information Governance Catalog, you can also analyze and classify the data governed by IBM StoredIQ based on the data classes that are synchronized from the governance catalog.

IBM StoredIQ Insights

IBM StoredIQ Insights provides dynamic and interactive filtering for your data with easy access to all metadata and instant plain-text preview of document content for full-text indexed volumes.

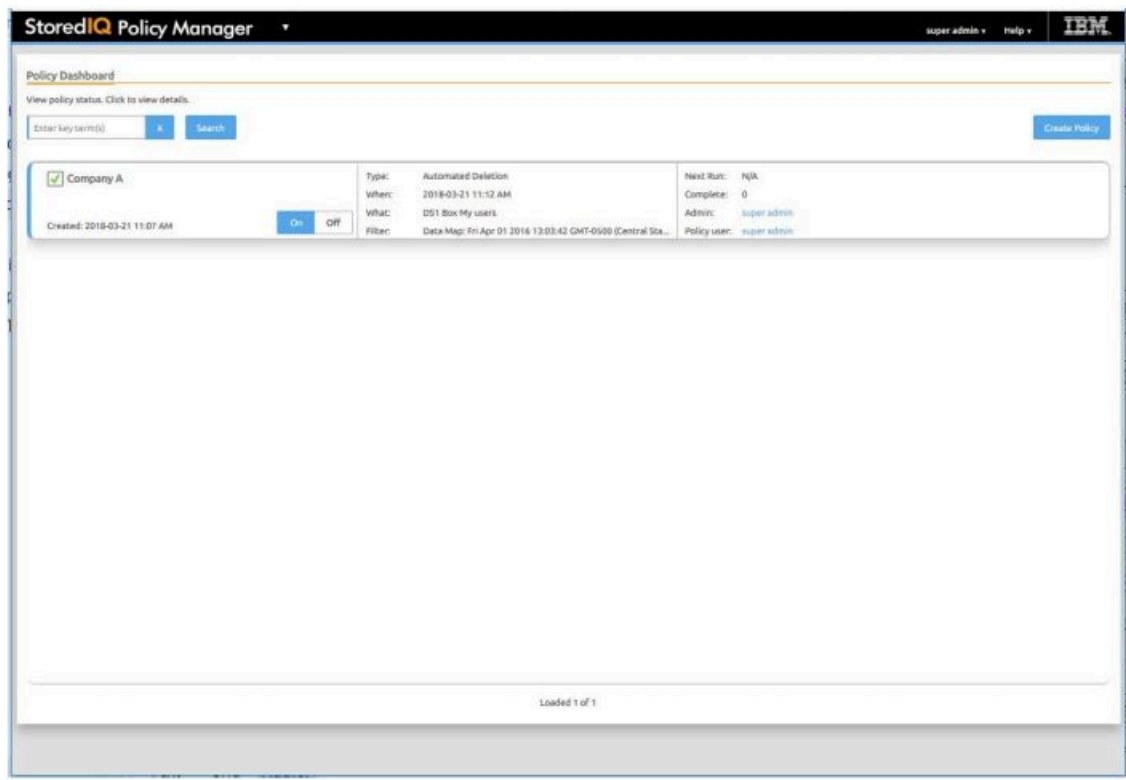
Faceted search lets you drill down to refine your search results as needed. In addition, you can apply any valid IBM StoredIQ filter query. Tags let you categorize the data for easier management. Visual representations of search results help you gain further insights into your data. Several chart types let you look at and explore data from different perspectives, thus helping you identify patterns and relationships very quickly.

With IBM StoredIQ Insights, you can search data that is managed and indexed by a data server of the type DataServer - Distributed. In mixed deployments that have classic and distributed data servers, only the content from distributed data servers will be searchable.



IBM StoredIQ Policy Manager

IBM StoredIQ Policy Manager allows users to run mature policies and processes at scale across a wider range of data.



The users can define and run systemwide policies, focusing on the execution of the process rather than understanding or reviewing affected data objects. Additionally, with reports of IBM StoredIQ Policy Manager, you can record what actions were conducted, when they were conducted, and what data was affected by the policy's execution.

IBM StoredIQ Desktop Data Collector

IBM StoredIQ Desktop Data Collector (also referred to as desktop client indexes desktops as volumes. The volumes appear in IBM StoredIQ Data Server and in IBM StoredIQ Administrator, where they can be used like any other data source.

The data server maintains an index using the information sent by the desktop client. After indexing, desktops - even offline or unreachable ones - can be viewed, searched, or targeted for later policy action.

Chapter 4. Planning for deployment

Planning for Deployment

When you plan a deployment of IBM StoredIQ, evaluate several infrastructure considerations.

In addition to the information in this section, review the requirements detailed in the IBM Software Product Compatibility Reports (SPCR) tool at: [Software Product Compatibility Reports: StoredIQ 7.6](#)

If you plan to use IBM StoredIQ for Legal Identification and Collection to create and manage data boxes and data requests that are to be fulfilled by IBM StoredIQ, also check the system requirements for StoredIQ for Legal at: [Software Product Compatibility Reports: StoredIQ for Legal 2.0.3](#)

Generate customized reports with the SPCR tool

Go to the page at [Software Product Compatibility Reports](#) to create a high-level report for supported operating systems, related software, hypervisors, and supported translations for any product. You can also create an in-depth report to get detailed system requirements, hardware requirements, and end of service information for each product. You can search for a product in all of the report types and reports are generated based on your query values.

The following report types are the most commonly generated reports from software product compatibility reports:

Detailed system requirements

When you select your product version for the detailed system requirements report, you can set a report filter for **Operating system platforms**, **Product components**, and **Capabilities**, including prerequisites and support software. After you view the report, you can save it as a URL to generate anytime or download it as a PDF.

Hardware requirements

When you select your product version for the hardware requirements report, you can set a report filter by the **Operating system families** option. Set the operating system filter by selecting some or all of the operating systems that are supported by your product. After you view the report, you can save it as a URL to generate anytime or download it as a PDF.

End of service

The end of service report shows the service window of the products that you specify over an eightyear span. For example, you can find out when your product is scheduled to go out of service.

Open Virtual Appliance (OVA) configuration requirements

Application stack

- vCPU:2
- Memory: 4 GB
- Storage:
 - Primary disk (vmdisk1): 21 GB
 - Data disk (vmdisk2): 10 GB

Gateway server

- vCPU:4
- Memory: 8 GB
- Storage:
 - Primary disk (vmdisk1): 100 GB
 - Data disk (vmdisk2): 75 GB
 - Swap disk (vmdisk3): 40 - 100 GB

Datasever - Classic

- vCPU:4

Even though increasing the number of vCPUs increases performance, the actual benefits depend on whether the specific host is oversubscribed or not.

- Memory: 16 GB

While the minimum value works under light-load condition, as the load increases, the data server quickly starts using swap space. For high load situations, increasing RAM beyond 16 GB can benefit performance. Monitoring swap usage can provide insight.

- Storage:
 - Primary disk (vmdisk1, SCSI 0:0): Default is 150 GB

This virtual disk has an associated VMDK that contains the IBM StoredIQ operating code. Do not change its size.



CAUTION

Attention

If you delete the primary disk, you delete the operating system, and the IBM StoredIQ software; the virtual machine might need to be scrapped.

- Data disk (vmdisk2, SCSI 0:1): Default is 2 TB

This virtual disk can be resized according to expectations on the amount of harvest data to be stored. For purposes of estimation, the index storage requirement for metadata is

about 30 GB per TB of managed source data. Full-text indexing requires an extra 170 GB per TB. Therefore, the default data disk size is targeted for managing 10 TB of source information.

- Swap disk (vmdisk3, SCSI 0:2): Default is 40 GB

When under load, the data server can use much RAM; therefore, having ample swap space is prudent. The minimum swap size is equal to the amount of RAM configured for the virtual machine. For best performance under load, place this disk on the highest speed data store available to the host.

The general size limits for a data server are 150 million objects or 500 defined volumes, whichever limit is reached first. Assuming an average object size of 200 KB equals about 30 TB of managed storage across 30 volumes of 5 million objects each, the index storage requirement for metadata on ~30 TB of storage that contains uncompressed general office documents is ~330 GB (11 GB per TB). Add 100 GB per TB of managed storage for full-text or snippet index. For example, to support 30 TB of storage that is indexed for metadata, you need 8 TB indexed for full-text search and extracted text (snippet cache) of 8 TB for auto-classification. A total of 1.9 TB of storage is required (metadata: 330 GB, full-text: 800 GB, snippet cache: 800 GB).

Data-server performance is impacted by the IOPS available from the storage subsystem. For each data server under maximum workload, at least 650 IOPS generally delivers acceptable performance. In the situations where there is a high load on the system, the IOPS that is used can reach up to 7000 with main write operations.

DataServer - Distributed

- vCPU:4
- Memory: 16 GB
- Storage:
 - Primary disk (vmdisk1, SCSI 0:0): Default is 150 GB
 - Data disk (vmdisk2, SCSI 0:1): Default is 2 TB
 - Swap disk (vmdisk3): (vmdisk3, SCSI 0:2): Default is 40 GB

If you deploy this type of data server, you must also deploy an Elasticsearch cluster with at least one node. If you deploy a cluster with more nodes, each of the Elasticsearch nodes must meet the listed requirements.

Each Elasticsearch node

- vCPU:1
- Memory: 32 GB
- Storage:
 - Primary disk (vmdisk1): 100 GB
 - Data disk (vmdisk2): 1 TB

The required memory depends on the index size that is handled by the data node. For a better performance, consider increasing the memory. The total memory available to the node must be distributed between the host operating system, the Elasticsearch java heap

space, and the kernel file system caches. For example, if the system has 32 GB memory, 2 GB must be allocated for the host operating system, 15 GB for the java heap space, and the remaining 15 GB for the file system caches.

The data disk (vmdisk2) can be resized according to expectations on the amount of harvest data to be stored.

Network and port requirements

For proper communication, the IBM StoredIQ components must be able to connect to each other.

You must enable network connectivity from the following locations:

- The data server IP address to the gateway IP address on port 11103
- The gateway IP address to and from the application stack IP address on port 8765 and 5432
- Ports 80, 443, and 22 from the administrative workstation to the application stack and data server IP addresses
- Port 22 from the administrative workstation to the gateway IP address.

TCP: port ranges for the firewall

To ensure the network access for desktop volumes, the following port ranges must be open through a firewall.

- 21000-21004
- 21100-21101
- 21110-21130
- 21200-21204

Default open ports

The following ports are open by default on the IBM StoredIQ.

SSH port 22

By default, port 22 is open on all IBM StoredIQ hosts. The port is used for Secure Shell (SSH) communication and allows remote administration access to the VM. In general, traffic is encrypted using password authentication. To add a layer of security, you can establish key-based authentication for passwordless SSH logins to any of the IBM StoredIQ nodes in your environment as described in [“Configuring SSH key-based authentication” on page 45](#).

Default open ports on the AppStack

Port number	Protocol
22	tcp
80	tcp

Port number	Protocol
443	tcp

Default open ports on the IBM StoredIQ data server

Port number	Protocol	Service
22	tcp	PROD-ssh
80	tcp	PROD-web
443	tcp	PROD-https (UI and Web Services APIs)
11103	tcp	PROD-transport (IBM StoredIQ transport services; communication between the gateway and the data server)
11104		

Enable or disable ports or services on the IBM StoredIQ data server

Enable or disable ports or services on the IBM StoredIQ data server

```
python /usr/local/storediq/bin/util/port_handler.pyc -parameter
```

-s

To list the current rules in iptables

-l

To list the supported services

-d port_number|'port_range'

To delete a port or a range of port numbers from iptables, for example:

```
python /usr/local/storediq/bin/util/port_handler.pyc -d '21200:21299'
```

-e 'service_name'

To enable a specific service, for example, to enable HTTPS services:

```
python /usr/local/storediq/bin/util/port_handler.pyc -e 'PROD-https'
```

-d 'service_name'

To disable a specific service, for example, to disable HTTPS services:

```
python /usr/local/storediq/bin/util/port_handler.pyc -d 'PROD-https'
```

Default open ports on the nodes in the Elasticsearch cluster

Port number	Protocol	Service
21	tcp	ftp
22	tcp	sshd
80	tcp	
443	tcp	
8888	tcp	SimpleHTTPServer (used for copying the siqelasticsearch.yml configuration file from the Elasticsearch node to the data server)
9200	tcp6	docker-proxy (listening for REST requests) You can restrict access to this port by either enabling Search Guard or by setting up a firewall. For more information, see “Securing Elasticsearch cluster communication with Search Guard” on page 51 or Restricting access to port 9200 on Elasticsearch nodes” on page 52.
9300	tcp6	docker-proxy (internode communication)

Default open ports on the IBM StoredIQ gateway

Port number	Protocol	Service
22	tcp	PROD-ssh
80	tcp	PROD-web
443	tcp	PROD-https (UI and Web Services APIs)
5432	tcp	PROD-postgres
5434	tcp	PROD-transport (IBM StoredIQ transport services; communication between the gateway and the data server)
8765		
7766		
11102		
11103		
11104		

Supported chain and rules on the IBM StoredIQ gateway

In iptables, the following firewall and chain rules are defined:

```
'PROD-transport': ['5434', '8765', '7766', '11102', '11103', '11104'],
      'PROD-https': ['443'],
      'PROD-ssh': ['22'],
      'PROD-web': ['80'],
      'PROD-postgres': ['5432']
```

```
'desktop' service:
      'PROD-broker': ['21000'],
      'PROD-collectionsvc': ['21300:21399'],
      'PROD-desktopupgrade': ['21004'],
      'PROD-objlistmgr': ['21100:21199'],
      'PROD-objlistsvc': ['21200:21299'],
      'PROD-registration': ['21001'],
      'PROD-session': ['21002'],
      'PROD-task': ['21003'],
```

Open ports for desktop client access to the data server

To open ports for desktop client access to the data server on OVA deployed systems, follow these steps:

1. Log in to the data server as root and run this command:

```
python /usr/local/storediq/bin/util/port_handler.pyc -e desktop
```

2. Run this command: `iptables -L INPUT`

In the output of the command, check the list position of the rule that is named PROD-reject, for example, the 6th position on the list.

3. Run this command: `iptables -A INPUT -j PROD-reject`

4. Run this command: `iptables -D INPUT list_position`

`list_position` is the position number of the PROD-reject rule that you determined in step 2.

5. Run the following command:

```
python /usr/local/storediq/bin/util/port_handler.pyc -e desktop
```

 **Tip:** These steps are required only on an IBM StoredIQ OVA deployed system. The correct ports are open on an upgraded system.

Environment sizing guidelines

To size an environment precisely, you must understand the factors such as harvest frequency, complexity of the source, and use case scenarios that drive application use and action execution.

The general design guidelines for IBM StoredIQ are as follows:

- For data servers of the type DataServer - Classic:
 - One data server for up to 30 TBs of data (which can vary based on the number of volumes, objects per volume, and object types).
 - Up to 500 volumes per data server.

Tip: When you're sizing an environment that includes Sharepoint data sources, keep in mind that volumes must be defined at Sharepoint site collection level, not the Sharepoint server level.

 - Up to 150 million objects per data server.
- One gateway per 50 data servers.
- One application server.
- NFS is slightly faster than CIFS for metadata only, but assume CIFS/NFS even for this exercise.
- Full-content processing of file (for example, .ZIP, .RAR, .GZ) and email archive (.PST, .NSF, .EMX) processing are slower as items must be extracted from the archives. If there is a significant number of these files in the file system and they are not excluded from content processing, the full-content processing rate can be too high. Until you have an initial index of the file system, you do not know how to weigh full-content processing of archives.
- An object/time metric is appropriate for metadata only NOT computing a hash, membership in the National Institute of Standards and Technology (NIST) or enumerating objects that are contained in archives. Converting it to a bytes/time metric is a function of the average object size and might vary tremendously. An average object size of 250 KB was used for the metric that is provided earlier.
- A bytes/time metric is appropriate for metadata-only computing a hash and full-content processing. The object per second rate can vary tremendously depending on the object type and sizes encountered. For example, processing an email or file archive is much more expensive than a PDF document.
- Metadata-only not computing a hash, membership in the NIST list, or enumerating objects that are contained in archives is requesting only the file-attribute information from the NAS. Individual files are not opened and read. The processing rate is high, but that does not translate into a large amount of data that traverses a network between the NAS and data server. The bytes/time rate does not translate into bytes served by the NAS and sent over the network.
- Metadata-only computing a hash, membership in the NIST list, or enumerating objects that are contained in archives opens and reads the contents of each file. The content of all requested files traverses the network between the NAS and data server. The maximum load that the data server can place on a NAS is metadata-only processing. It requires all file content to be read to compute a hash or enumerate objects that are contained in archives. The bytes/time rate translates into bytes served up by the NAS and network traffic that must be considered.

- Full-content processing opens and reads the contents of each file to extract all text. The content of all requested files traverses the network between the NAS and data server. The processing time to enumerate archives, extract text, index words, and extract entities on the data server reduces the rate that data is requested from a NAS compared to metadata-only with full hash. The bytes/time rate translates into bytes served up by the NAS and network traffic that must be considered.
- The interrogator process count on the data server for "metadata only not reading all content indexing" is set to eight for optimal performance.
- The interrogator process count for all other processing that involves reading all content default setting is four per data server.
- The interrogator count can be viewed as the number of client connections that are made to a data source actively requesting data. It is important for capacity planning for the data source.
- The data servers are assumed to be "network close" to the NAS data sources. Network latency under 10 ms with at least 1000 Mbps bandwidth is assumed (connected through local area network). The data servers need a low latency high-bandwidth connection to a NAS data source for acceptable indexing performance.
- The gateway and application stack can be located remotely from the data servers. Network connections with latency greater than 10 ms and bandwidth of at least 2+ Mbps are acceptable.

VMware requirements

- VMware vSphere v5.0 and fix packs or v6.0 and fix packs.
- VMware ESXi v5.0 and fix packs, v6.0 and fix packs, or v6.5 and fix packs.
- VMware virtual machine hardware version 8.0 or later. For more information, see the [VMware product documentation](#).
- The appropriate VMware license to enable the required processor cores and memory for the virtual machine.

Stack-provisioning prerequisites

Before a deployment, verify that you meet these prerequisites.

- At least one physical server with sufficient processor, RAM, and hard disk configuration for the planned management project.
- VMware ESX or ESXi on CD/DVD or USB drive.
- IP addresses, cables, and physical switch ports for at least the ESXi/ESX interface, one data server, one gateway server, and one application stack.
- Network connectivity that is enabled from the following locations:
 - The data server IP address to the gateway IP address on port 11103
 - The gateway IP address to and from the application stack IP address on port 8765 and 5432
 - Ports 80, 443, and 22 from the administrative workstation to the application stack and data server IP addresses

- Port 22 from the administrative workstation to the gateway IP address.
- Network connectivity that is enabled from the data server IP address to any data sources to be harvested and managed.
- A management station computer or notebook from where the load-management work is done.

License usage metrics

Using the IBM License Metric Tool, you can generate license consumption reports that count IBM StoredIQ license usage.

IBM StoredIQ is licensed by Resource Value Unit (RVU). RVU calculation is based on terabytes IBM StoredIQ.

On the IBM StoredIQ application stack, a license program writes usage information to an IBM Software Licence Metric Tag (SLMT) file. This file has the extension .slmtag and can be read periodically by the IBM License Metric Tool (ILMT) after it has been configured to scan for these files. You can generate reports that summarize usage.

By default, the license program retrieves the size of the All Data Objects info set in terabytes once per day and writes this information to the /var/siq/ilmt/3cd1469042433ee7010fe09f661dc67b.slmtag file. The .slmtag file can store information up to a maximum file size of 1 MB, after which the file is archived and a new log file is created. A maximum of 10 log files are kept.

The .slmtag file contains usage information in the following format, where new metric records are appended to the end of the file:

```
<SchemaVersion>2.1.1</SchemaVersion>
  <SoftwareIdentity>
    <PersistentId>cb98e260a2a14872902578de1b8e2016</
PersistentId>
    <Name>IBM StoredIQ Data Assessment</Name>
    <InstanceId>/var/siq/ilmt</InstanceId>
  </SoftwareIdentity>
  <Metric logTime="2019-01-25T15:58:26+00:00">
    <Type>TERABYTE</Type>
    <SubType>All Data Objects</SubType>
    <Value>0.005</Value>
    <Period>
      <StartTime>2019-01-25T15:58:26+00:00</StartTime>
      <EndTime>2019-01-25T15:58:26+00:00</EndTime>
    </Period>
  </Metric>
```

Integration with IBM License Metric Tool

Versions of IBM License Metric Tool (ILMT) that support IBM Software License Metric Tag (SLMT) can generate license consumption reports. An ILMT agent can scan in configurable intervals the file system for .slmtag files, collect information, and send it to the corresponding ILMT server.

ILMT reports the number of terabytes managed by IBM StoredIQ. This number is to be used as input for the RVU License Conversion Table specified in the license information (li_languagecode file) that comes with IBM StoredIQ. On the application stack, you can find the license information in the License directory.

For more information about using IBM License Metric Tool, see the IBM License Metric Tool documentation.

Security

Plan and implement specific security measures to protect the application and the data it manages, especially when you deploy IBM StoredIQ into sensitive environments.

IBM StoredIQ keeps your data secure through encryption, security hardening, and auditing.

Federal Information Processing Standard (FIPS)

FIPS is a standard recommended by the National Institute of Standards and Technology (NIST) and the US Federal Government. It ensures certain security standards are met for software or hardware components deployed at US government sites. Enabling FIPS ensures that the SSL/TLS engine that is compliant with the US Government recommendation is used. IBM StoredIQ supports FIPS Level 1.

Secure gateway communication can be enabled without FIPS. If FIPS is enabled, IBM StoredIQ uses FIPS compliant versions of OpenSSL.

Secure communication and encryption of data in motion

In a production environment, you should configure or install certificates on the AppStack to enable HTTPS communication and to enable encryption of data in motion between the browser and the AppStack. You can do this during installation and initial configuration or at any time afterward. For details, see the instructions for configuring certificates.

The gateway handles the communication between the data servers and the application stack. By default, the communication between the gateway, any data servers, and the AppStack is in plain text and is not encrypted. If your enterprise security policy mandates encryption of data in motion, enable secure gateway communication. In this case, secure gateway communication must be configured on all three IBM StoredIQ components. You can enable secure gateway communication during installation and initial configuration or at any time afterward. For details, see “Managing the status of secure gateway communication” on page 54.

IBM StoredIQ then uses stunnel to ensure secure communication between the components. If your environment includes data servers of the type DataServer - Distributed, stunnel can also be used to encrypt the communication between the nodes within the Elasticsearch cluster but not for encrypting the communication between the data server and the Elasticsearch cluster.

To secure the communication between the data server and the Elasticsearch cluster and the communication within the Elasticsearch cluster likewise, you can enable Search Guard. For more

information, see “Securing Elasticsearch cluster communication with Search Guard” on page 51. If you don't want to do that but still want to restrict client access to port 9200 on the Elasticsearch nodes, you can set up the firewall accordingly. For more information, see “Restricting access to port 9200 on Elasticsearch nodes” on page 52.

If FIPS is not enabled, the following cipher suites and encryption algorithm are used for data at rest:

TLS_ECDHE_ECDSA_WITH_AES_128_GCM_SHA256

TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256

You can configure these cipher suites in the configuration files listed in the list of key and certificate files. However, if you run the utilities for enabling stunnel, you might need to make the respective configuration changes again.

Encryption of data at rest

Starting with IBM StoredIQ version 7.6.0.15, the disk volume on which the Elasticsearch indexes are stored is encrypted by default. IBM StoredIQ uses Linux Unified Key Setup (LUKS) for disk encryption. For details about key management, see “Key and certificate management” on page 41.

Optionally, you can encrypt the application data on the IBM StoredIQ application stack. For more information, see “Enabling encryption of IBM StoredIQ AppStack application data” on page 50.

Network isolation

If full-text harvesting and Step-up Analytics actions (cartridges) are applied, Elasticsearch indexes can contain potentially sensitive content. Therefore, you should deploy the Elasticsearch nodes in an isolated location on the network (for example, as an enclave or behind a firewall) that is properly secured according to the sensitivity of the data being harvested. Only the IBM StoredIQ application stack and data servers should be allowed to communicate with the Elasticsearch nodes.

Also, any data servers and the gateway should be deployed in an isolated network location to allow for communication with authorized clients only.

Access control

The following administrative accounts are required. The builder and siqadmin accounts are IBM StoredIQ-specific accounts. For more information about these accounts, see “Default user accounts” on page 17 (*on page*).

root and builder accounts on the Elasticsearch cluster nodes

Remote login for root can be disabled. However, local root login is required, either log in as root or use the su command to obtain root permissions temporarily.

You set the passwords for the root and builder accounts during the configuration process when you start the VM for the first time. You can change these passwords anytime.

siqadmin account on the AppStack

Administration of the AppStack usually does not require direct root access. For day-to-day administration, the siqadmin account can be used.

You set the password for the siqadmin account during the configuration process when you start the VM for the first time. You can change this password anytime.

Chapter 5. Deploying IBM StoredIQ

Deploying IBM StoredIQ

IBM StoredIQ is a virtual appliance that you deploy and configure in a VMWare virtual host environment.

Deploying the virtual appliances

Use VMware vSphere Client to deploy the virtual appliances to an ESX server. Deploy OVAs for the gateway, the data server, and the application stack, preferably in this order. If an Elasticsearch cluster is also being deployed, those OVAs must be deployed first.


Ensure that all prerequisites for the deployment described in the planning section are met and that the required software packages are available on your local system before you start this task.

Download the latest version of IBM StoredIQ from either IBM Fix Central or IBM Passport Advantage®. For information about the package names and part numbers and the links to the proper download locations, see the [download document](#).

The number of Elasticsearch OVAs deployed depends on the planned size of your Elasticsearch cluster. The default setup is a three-node cluster. Each Elasticsearch node requires a separate OVA deployment.

The number of data server OVAs deployed depends on the number of data servers needed. Each data server requires a separate OVA deployment.

1. Connect to the ESX server or vCenter server.
2. Open VMware vSphere Client.
3. Select **File > Deploy OVF Template**.
4. Within the **Deploy OVF Template** wizard, complete these steps.
 - a. Within the **Select source** page, click **Local file**, and then browse to and select the appropriate OVF. Click **Next**
 - b. Within the **Review details** page, review the OVF template details. These storage requirements are critical and will be used to select a data store during deployment. Click **Next** to proceed.
 - c. Within the **Select name and folder** page, enter a name for the deployed template or use the default name. Click **Next**.

- d. Within the **Select a resource** page, select the resource pool where the deployed OVF template runs. Click **Next**.
 - e. Within the Select storage page, select a data store on which to store the deployed OVF template files. Click Next.
 - f. Within the Disk Format list, select the disk format to be deployed. Note that although Thin Provision saves disk space, it can negatively affect performance. If possible, select Thick Provision Lazy Zeroed. Click Next.
 - g. Within the Network Mapping, map the network to the appropriate network. Click Next.
 - h. Within the Ready to Complete page, review the deployment settings. Click Finish to complete the Deploy OVF Template.
At this point, you can select the Power on check box to turn on the VM after deployment is complete.
-  **Important:** If your IBM StoredIQ environment includes an Elasticsearch cluster, do not select this option when you're deploying the data server OVA. The Elasticsearch cluster setup must be complete before you run the data server first-boot process.
5. Repeat steps 3 and 4 for each OVA.

Configuration

Configure the components of your IBM StoredIQ environment in this order:

1. Elasticsearch cluster (if applicable)
2. Gateway
3. One or more data servers
4. Application stack

Deploying IBM StoredIQ on Microsoft Hyper-V

As an alternative to installing on an ESX server, the IBM StoredIQ gateway, data server, and application stack can be installed on Microsoft Hyper-V. This option is not supported for the Elasticsearch virtual appliance.

Installing IBM StoredIQ on Microsoft Hyper-V requires the use of third-party software. For the procedure described here, the following software prerequisites must be met:

- Microsoft Hyper-V Manager must be installed on a Windows system.

- 7-zip for Windows or tar for the Linux operating system must be available to extract contents of the OVA.
- 7-zip for Windows or gunzip for the Linux operating system must be installed to decompress the vmdk.gz file.
- Microsoft Virtual Machine Converter 3.0 (Windows) or qemu-img (Linux) or must be installed to convert the files from VMWare .vmdk files to Hyper-V .vhdx files.

sample procedure

Microsoft Hyper-V does not support OVAs and OVF. The virtual machine needs to be built manually.

Consider the following instructions a sample procedure. The instructions might differ depending on the version of the third-party software.

Complete these steps for the gateway, the data server, and the AppStack:

1. Extract the vmdk file from the OVA.
 - a. Download the OVA.
 - b. Extract the contents of the OVA by using 7-zip for Windows or tar for the Linux operating system.
 - c. Decompress the vmdk.gz archive by using 7-zip for Windows or gunzip for Linux.
 - d. Delete everything except the vmdk file when the OVA extraction is complete.
2. Convert the VMWare .vmdk file to a Microsoft Hyper-V .vhdx file.

For instructions about converting .vmdk files to .vhdx files, see the following links:

- Use [Microsoft Virtual Machine Converter](#)
- [Using qemu-img](#)

3. Build a virtual machine.
 - a. Select **New Virtual Machine > Next** from Hyper-V Manager.
 - b. Enter the name of the virtual machine in the **Specify Name and Location** window and click **Next**.
 - c. Select **Generation of the machine** in the **Specify Generation** window.
 - d. Enter the startup memory in the **Assign Memory** window.
 - e. Select the correct network in the **Configure Network** window and click **Next**.
 - f. Select **Use an existing virtual hard disk** in the **Connect Virtual Hard Disk** window. Search the disk location, select one of the primary disks, and then click **Next**.
 - g. Click **Finish**.
 - h. Repeat these previous steps for all IBM StoredIQ disks.
4. Optional: Add disks
 - a. Select a virtual machine in Microsoft Hyper-V Manager.
 - b. Select the settings for the virtual machine under **Actions**.

- c. Select the IDE controller and highlight Hard Drive in the **Settings** window.
 - d. Select Hard Drive and virtual hard disk and click **New**.
 - e. Click **Next** in the New Virtual Hard Disk wizard.
 - f. Select **VHDX** in Choose Disk Format and then click **Next**.
 - g. Select **Dynamically expanding** in Choose Disk Type and then click **Next**.
 - h. Enter a name for the disk in the **Specify Name and Location** window and then click **Next**.
 - i. Select **Create a new blank virtual hard disk** and enter a size in the **Configure Disk** window and then click **Finish**.
 - j. Repeat this procedure for all additional disks.
5. Complete the installation process by following the instructions in “Configuring IBM StoredIQ” on page 22.

Complete the installation process by following the instructions in “Configuring IBM StoredIQ” on page 22 (*on page*).

Chapter 6. Information

Chapter 7. Glossary

Glossary Terms

Application stack

The application stack provides the user interface for the IBM StoredIQ Administrator, IBM StoredIQ Data Workbench, IBM StoredIQ Insights, and the IBM StoredIQ Policy Manager products.

The synchronization feature for integration with a governance catalog is also part of the application stack.

Cartridges

Cartridges are compressed files that contain analysis logic. When you add a cartridge to IBM StoredIQ AppStack, it can detect new data in documents during indexing and make these new insights searchable. For example, a sensitive pattern cartridge can enable IBM StoredIQ to detect passport numbers, phone numbers, and other IDs.

To apply the analysis logic contained in the cartridge, you must run a Step-up Analytics action that uses the cartridge on an info set. IBM StoredIQ examines all documents in the info set, applies the analytics, and then stores the analysis results in the IBM StoredIQ index.

Connector API SDK

A connector is a software component of IBM StoredIQ that is used to connect to a data source such as a network file system and access its data. Using IBM StoredIQ Connector API SDK, developers of other companies can develop connectors to new data sources outside the IBM StoredIQ development environment. These connectors can be integrated with a live IBM StoredIQ application to index, search, manage, and analyze data on the data source.

Data servers

A data server obtains the data from supported data sources and indexes it. By indexing this data, you gain information about unstructured data such as file size, file data types, file owners.

The data server pushes the information about volumes and indexes to the gateway so it can be communicated to the application stack. Multiple data servers feed into a single gateway.

Data servers can be categorized in two types: DataServer - Classic and DataServer - Distributed. With a data server of the type DataServer - Distributed, the index is stored in an Elasticsearch cluster.

Data servers of this type also provide better performance in search queries. They can manage much larger amounts of data than data servers of the type DataServer - Classic, thus making the IBM StoredIQ deployments more scalable.

DataServer - Classic

Data servers can be categorized in two types: DataServer - Classic and DataServer - Distributed. DataServer - Classic refers to the regular data servers. It uses either the current PostgreSQL or Lucene index as an index.

DataServer - Distributed

The distributed data server uses an Elasticsearch cluster instead of an embedded Postgres database. It increases the scalability and flexibility of the IBM StoredIQ deployment in a way that it can manage much larger amounts of data. Without adding more data servers, data that is managed by the IBM StoredIQ deployment can be increased by adding new nodes to the Elasticsearch cluster. Search queries perform better on DataServer - Distributed.

Gateway

The gateway communicates between the data servers and the application stack. The application stack polls the gateway for information about the data on the data servers. The data servers push the information to the gateway.

IBM StoredIQ Insights

IBM StoredIQ Insights provides dynamic and interactive filtering for your data with easy access to all metadata and instant plain-text preview of document content for full-text indexed volumes.

Faceted search lets you drill down to refine your search results as needed. In addition, you can apply any valid IBM StoredIQ filter query. Tags let you categorize the data for easier management. Visual representations of search results help you gain further insights into your data. Several chart types let you look at and explore data from different perspectives, thus helping you identify patterns and relationships very quickly.

With IBM StoredIQ Insights, you can search data that is managed and indexed by a data server of the type DataServer - Distributed. In mixed deployments that have classic and distributed data servers, only the content from distributed data servers will be searchable.

Index

A

7,21

- application stack
- see* AppStack
- auto-classification models
- see* automated document classification

D

- Data servers
 - DataSeter - Classic
 - 7,21
 - DataSeter- Distributed
 - 7,21
- Data Workbench
 - about
 - 7,21
 - potential uses of
 - 7,21
- deploy
 - OVA
 - 7,21
 - OVF
 - 7,21
 - virtual appliance
 - 7,21

I

- IBM Licence Tool Metric
 - see* ILMT
- IBM StoredIQ Desktop Data Collector
 - see* desktop client

L

- licenced programs
 - descriptions
 - 7,21

O

- Open Virtual Appliance
 - see* OVA

R

- Resource Value Unit
 - see* RVU

T

- TCP
 - port ranges