

IBM StoredIQ

*Overview Guide*



**Note**

Before using this information and the product it supports, read the information in [Notices](#).

This edition applies to Version 7.6.0.22 of product number 5724M86 and to all subsequent releases and modifications until otherwise indicated in new editions.

© **Copyright International Business Machines Corporation 2001, 2020.**

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

---

# Contents

- IBM StoredIQ product library..... iv**
- Contacting IBM StoredIQ customer support.....v**
  - Contacting IBM..... v
- Overview of IBM StoredIQ..... 1**
  - Solution components..... 2
- Applications of IBM StoredIQ.....4**
  - IBM StoredIQ Data Server..... 4
  - IBM StoredIQ Administrator..... 5
  - IBM StoredIQ Data Workbench..... 6
  - IBM StoredIQ Insights..... 8
  - IBM StoredIQ Cognitive Data Assessment..... 8
  - IBM StoredIQ Desktop Data Collector..... 9
- User roles of IBM StoredIQ..... 10**
- Key terms.....11**
  - Volumes..... 11
  - IBM StoredIQ index..... 11
  - Information set..... 11
  - Filter..... 12
  - Overlay..... 12
  - Set Ops..... 12
  - Node ops..... 12
  - Duplicate operation..... 13
  - Enhancement..... 13
  - Action..... 13
  - Report..... 13
  - Exceptions..... 14
- Notices.....15**
  - Trademarks..... 16
  - Terms and conditions for product documentation..... 17
  - IBM Online Privacy Statement..... 17
- Index..... 19**

## IBM StoredIQ product library

---

The following documents are available in the IBM® StoredIQ® product library.

- *IBM StoredIQ Overview Guide*
- *IBM StoredIQ Deployment and Configuration Guide*
- *IBM StoredIQ Data Server Administration Guide*
- *IBM StoredIQ Administrator Administration Guide*
- *IBM StoredIQ Data Workbench User Guide*
- *IBM StoredIQ Cognitive Data Assessment User Guide*
- *IBM StoredIQ Insights User Guide*
- *IBM StoredIQ Integration Guide*

The most current version of the product documentation can always be found online: [https://www.ibm.com/support/knowledgecenter/en/SSSHEC\\_7.6.0/welcome/storediq.html](https://www.ibm.com/support/knowledgecenter/en/SSSHEC_7.6.0/welcome/storediq.html)

## 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 Center](#).

## 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/>.



---

# Overview of IBM StoredIQ

IBM StoredIQ provides scalable analysis and governance of unstructured data in-place across disparate and distributed email, file shares, desktops, and collaboration sites. Its products enable companies to discover, analyze, and act on data for eDiscovery; records retention and disposition; compliance; and storage optimization initiatives.

## **Powerful solutions for managing unstructured data in-place**

IBM StoredIQ addresses the problems that challenge records management, electronic discovery, compliance, storage optimization, and data migration initiatives. By providing an in-depth assessment of unstructured data where it is, IBM StoredIQ gives organizations visibility into data to make more informed business and legal decisions.

IBM StoredIQ delivers:

- In-place data management that allows an organization to discover, recognize, and act on unstructured data without moving it to a repository or specialty application.
- A powerful search function that accelerates the understanding of large amounts of unstructured content.
- Simplified analysis of large amounts of corporate data to provide detailed analysis faster and limit the impact on user productivity by analyzing and managing data in-place.
- Intelligence that supports many different policy actions such as copy, delete, move, copy to retention, or export.

## **An organized, systemic, and defensible approach to eDiscovery**

IBM StoredIQ provides insight into enterprise data to help ease the costs and efforts that are involved in electronic discovery (eDiscovery) response. IBM StoredIQ helps decrease the volume of unstructured data by targeting only the most relevant information to a particular case and providing forensically sound and defensible collections.

IBM StoredIQ delivers:

- Faster access to relevant information before collection, giving legal and IT teams the data that is needed to make more informed legal decisions.
- A powerful search function that accelerates the understanding of large amounts of unstructured content and encourages organizational alignment that can lead to reduced legal risks and costs.
- Simplified analysis of large amounts of corporate electronically stored information (ESI), providing faster detailed analysis, and limiting the impact on user productivity.
- Intelligence that allows companies to respond more quickly to litigation with the most relevant data.

## **Information governance to automate policy and compliance across unstructured data**

IBM StoredIQ helps organizations identify, classify, and manage enterprise information according to business value to reduce risk and cost. Corporations can gain a deeper and holistic understanding of their unstructured data to address business and regulatory requirements, compliance enforcement, data retention and respond to audit requests.

IBM StoredIQ provides the following features and solutions:

- A powerful data assessment solution for discovering, recognizing, and acting on unstructured data without first moving it to a repository.
- Advanced search capabilities that are tailored to help legal, records, compliance, and IT staff discover data in accordance with corporate and regulatory policy.
- Detailed data analysis to simplify the analysis of large amounts of corporate data.
- In-place data management capabilities to remediate regulatory and corporate policy violations.

### **Flexible solution for identifying and collecting data from remote devices**

IBM StoredIQ Desktop Data Collector enables organizations to apply corporate governance policies to user desktops and notebooks. Users can identify and collect corporate records or custodian data for legal matters.

Desktop Data Collector delivers:

- A powerful, flexible solution for identifying and collecting data for investigations, litigation matters, or records retention.
- A simplified collection of information on remote desktops and notebooks.
- Centralized management to minimize IT burden and improve efficiency.
- Intelligent desktop data collection for identifying corporate records or custodian data and collecting them to a central repository.

### **Bridging structured and unstructured content for enterprise information governance**

A unified governance architecture is the key to governing all enterprise content, be it structured or unstructured, on premises or in the cloud, thus helping organizations manage the findability, usability, and integrity of their data.

Integrating IBM StoredIQ with a governance catalog bridges structured and unstructured content to enable enterprise information governance.

## **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 IBM StoredIQ Cognitive Data Assessment 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.

The screenshot displays the IBM StoredIQ Data Server user interface. At the top, there is a navigation bar with the IBM logo, a 'DS Admin' button, and tabs for 'Administration', 'Folders', and 'Audit'. Below this is a secondary navigation bar with 'Dashboard', 'Data sources', and 'Configuration' tabs. The main content area is divided into several sections:

- Page refresh:** Off | 30 sec | 60 sec | 90 sec
- Today's job schedule:** No jobs scheduled for today.
- Jobs in progress:** No jobs are currently running.
- System summary:** View a summary of system details.

Total system data objects	10756
Total contained data objects	1081591
Total data objects	1092347
Number of volumes	13
Date of last completed harvest	No harvests run.
- Harvest statistics:** Review the performance over the last hour for all harvests.

Processes	4
Average data objects per second	0.0
Average data object size	0 bytes
Maximum data object size	0 bytes
Average data object processing time	0.0 sec
Maximum data object processing time	0.0 sec
- Event log:** The current event log as of 03/21/2018 05:03 PM. Includes links for 'Clear this view', 'Download today's event log', and 'View all event logs'. A list of 'Last 500 events' is shown with details like '[INFO][bmorgantrunkdemo-ds1][Mar 21, 2018 08:00:07]: Database compactor completed (41003). [Subscribe](#)'.
- Appliance status:** Shows a green status indicator and a 'Controller' button. Includes links for 'About appliance' and 'View cache details'.

# 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, infosets, 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.
- **Auto-classification:** Automated document categorization, what IBM StoredIQ refers to as auto-classification 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.
- **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.

The screenshot shows the IBM StoredIQ Data Workbench interface. At the top, there is a navigation bar with the IBM logo, 'My Requests', 'super admin', and 'Help'. Below this is the 'Infoset Dashboard' header, followed by a sub-header 'Infoset status and state. Click to view and create advanced infosets.' A search bar labeled 'Filter By Name:' is present with a search button and a 'Select' button. The main content is a table with the following columns: Name, Total objects, Infoset size, Composition, Created, Type, and Description. The table lists various data objects such as 'All Data Objects', 'All objects from SP (2010&...', 'All System-Level Objects', etc.

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 (2010&...	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 ds2	423	37.92 MB	Mixed Level	2016-03-29 9:25 AM	System	
big12 ds2 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 P8 nimmo8	58	3.28 MB	Mixed Level	2015-12-13 11:06 PM	User	

Loaded 49 of 49

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.

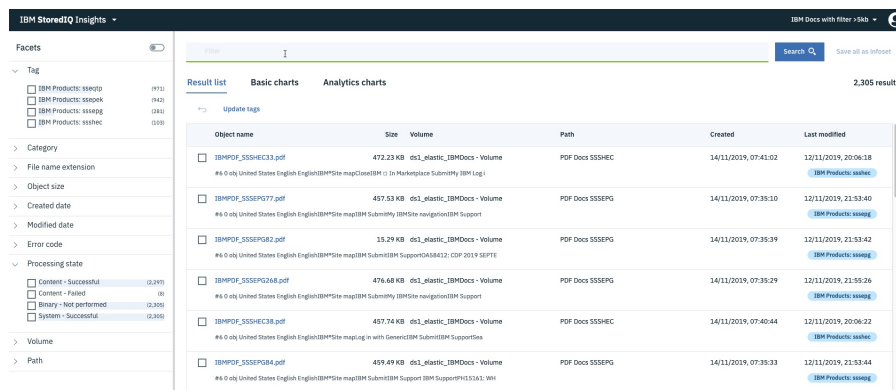
- You need to find all company email that is sent from or received by Eileen Sideways (esideways@thecompany.com). You can use IBM StoredIQ Data Workbench to find all email and then copy that data to a predefined repository. You can also use IBM StoredIQ Data Workbench to find all of the esideways@thecompany.com 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 infosets 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.



The screenshot displays the IBM StoredIQ Insights interface. On the left, there is a 'Facets' sidebar with various filters like Tag, Category, File name extension, Object size, Created date, Modified date, Error code, Processing state, and Volume. The main area shows a search results table with the following columns: Object name, Size, Volume, Path, Created, and Last modified. The table contains several rows of PDF documents, each with a checkbox, object name, size, volume, path, creation date, and last modified date. The results are filtered to show 2,305 results.

Object name	Size	Volume	Path	Created	Last modified
<input type="checkbox"/> IBMPOF_S5SHEC33.pdf	472.23 KB	os1_elastic IBMDocs - Volume	PDF Docs S5SHEC	14/11/2019, 07:41:02	12/11/2019, 20:06:18
<input type="checkbox"/> IBMPOF_S5SEPG77.pdf	457.53 KB	os1_elastic IBMDocs - Volume	PDF Docs S5SEPG	14/11/2019, 07:39:10	12/11/2019, 21:53:40
<input type="checkbox"/> IBMPOF_S5SEPG82.pdf	15.29 KB	os1_elastic IBMDocs - Volume	PDF Docs S5SEPG	14/11/2019, 07:39:39	12/11/2019, 21:53:42
<input type="checkbox"/> IBMPOF_S5SEPG84.pdf	476.68 KB	os1_elastic IBMDocs - Volume	PDF Docs S5SEPG	14/11/2019, 07:39:29	12/11/2019, 21:55:26
<input type="checkbox"/> IBMPOF_S5SHEC38.pdf	457.74 KB	os1_elastic IBMDocs - Volume	PDF Docs S5SHEC	14/11/2019, 07:40:44	12/11/2019, 20:06:22
<input type="checkbox"/> IBMPOF_S5SEPG84.pdf	459.49 KB	os1_elastic IBMDocs - Volume	PDF Docs S5SEPG	14/11/2019, 07:39:33	12/11/2019, 21:53:44

## IBM StoredIQ Cognitive Data Assessment

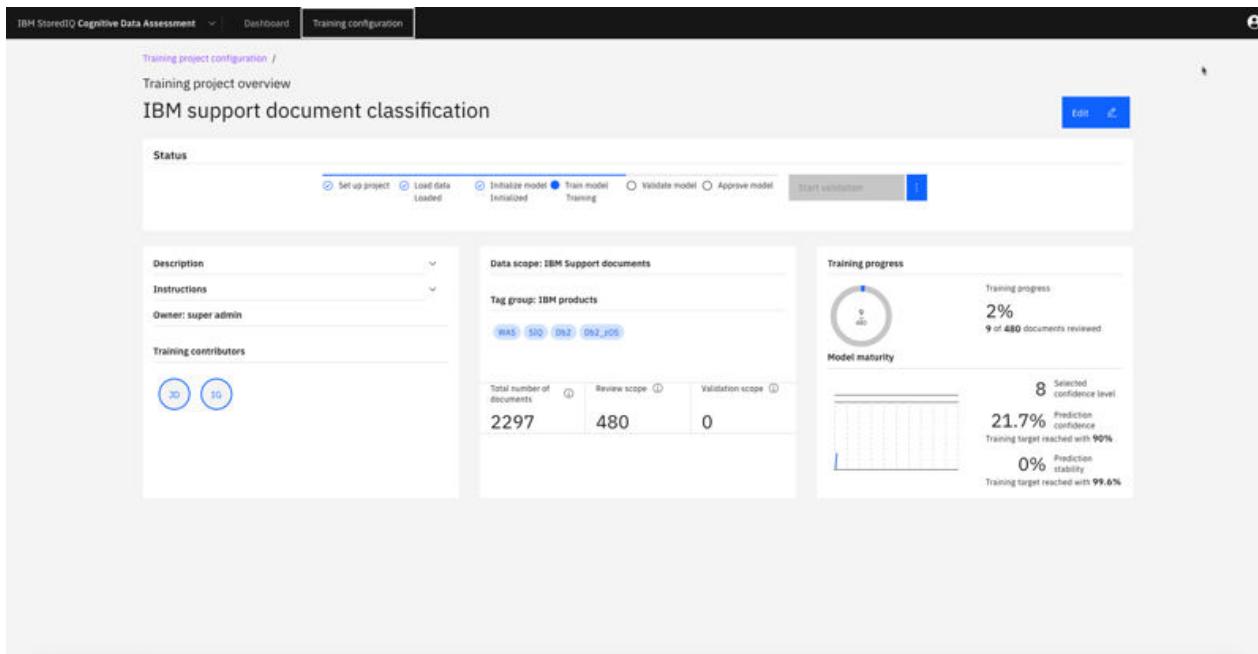
With IBM StoredIQ Cognitive Data Assessment, your organization can vastly improve the efficiency, accuracy, and automation of document classification decisions.

Gaining actionable insight in your unstructured data most often requires assessing and reviewing documents, no matter what the use case is:

- e-discovery
- Data cleanup
- Compliance and audit activities
- Retention
- Sensitive data management

To categorize your data properly, unstructured documents of various formats and different length must be classified or tagged. To minimize the time and effort spent on tagging, you can create a machine-learning model by using IBM StoredIQ Cognitive Data Assessment.

Cognitive Data Assessment streamlines the creation of a model. It combines the training and validation of the model where users contribute to the process in a training project by accepting or rejecting the suggested classification. After the model is built, it can automatically tag new documents for you. When the model is deemed mature and is approved, it can be downloaded and deployed as a cartridge and applied to any IBM StoredIQ infoset. The classifications are then readily available in IBM StoredIQ Insights.



## 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 searched and acted upon.

---

# User roles of IBM StoredIQ

IBM StoredIQ provides applications and interfaces for four user types.

## **System administrator**

This person is responsible for IBM StoredIQ installation setup, system and network configuration and maintenance, and administration activities. These activities are required to be done before other users of the IBM StoredIQ applications can start their work. The administrative activities mainly include:

- Getting the data servers and data centers ready for use.
- Adding volumes and ensuring security of data and data source.
- Harvesting volumes and generating system info sets.
- Managing users, actions, and target sets.
- Managing cartridges.
- Creating reports and using auto-classification models.
- Setting up the integration with a governance catalog.

For more information about what system administrators do, see the administration information.

## **Data expert**

A data expert understands both business processes and technical implementation. This person is responsible for responding to requests from the business user in a timely fashion, assessing and managing data in the IBM StoredIQ applications, such as IBM StoredIQ Data Workbench and IBM StoredIQ Insights. This person also decides which data is published to the governance catalog.

For more information about what a data expert does, see the information about managing your data.

## **Cognitive Data Assessment users**

A Cognitive Data Assessment user can either be a project owner or a project contributor. A project owner is responsible for setting up and managing a training project for creating a CDA classification model. A project contributor helps train and validate a model by reviewing the predictions.



---

## Key terms

The following terms are key to understanding IBM StoredIQ as a whole.

### Volumes

---

A volume represents a data source or destination that is available on the network to IBM StoredIQ.

Within IBM StoredIQ, these volume types exist:

**Primary**

The storage where the unstructured content resides (*data source*) and is harvested from.

**Retention**

Storage for information to be retained for a set amount of time or for litigation hold. Typically, a retention volume is immutable.

**Export**

Storage to keep the data produced from a policy so that it can be exported as a load file and uploaded into a legal review tool. Administrators can also configure export volumes for managing harvest results from cycles of a discovery export policy.

**System**

Storage for data that is used for application specific purposes.

### IBM StoredIQ index

---

Based on the volumes that are added in the data servers, indexes are generated through IBM StoredIQ Administrator or IBM StoredIQ Data Server to examine, classify, and map data in the network. They are used to search for data and find out what and how much data you have in your system.

There are two types of indexes: metadata index and full-text index.

Metadata index contains all the information about the data at a specific location on a network. It includes descriptive information or attributes about the data such as a file name, file size, created date, and owner.

Full-text index is a more detailed index on the contents of the data itself. By reading the contents of the data, the words or characters that are contained within the data can be referred to and searched against.

### Information set

---

An information set, abbreviated as an info set, is the core concept in using the IBM StoredIQ applications. It is created and used to collect specific data to manage the business system.

There are two types of info sets: system info sets and user info sets.

**System info set**

These info sets are generated after volumes are harvested from IBM StoredIQ Administrator or IBM StoredIQ Data Server. They can be viewed from the IBM StoredIQ Data Workbench user interface but users cannot edit or delete them. They can also be manually created by the administrator to target certain volumes in the IBM StoredIQ Administrator application. These system info sets can be edited or deleted by the system administrator. System info sets must be generated or created before any user info sets can be created.

**User info set**

A user info set is created out of an existing system info set by a user. It is defined to contain specific data that a user needs to operate upon the system. For example, you can create a user info set that contains a person's emails in Year 2000. Then, you can act on the data within this info set: to move, copy, or delete it from your system.

## Data Map

Data Map is the visualization of an infoset. It provides a visual layout of the data and in-depth information about data source types, data categories, size or amounts, the number of data objects and details. For more information, see the Data Workbench guide.

## Filter

---

A filter is created upon the available information that was populated by the index. It is used to classify or refine the existing infoset to create a new infoset.

A filter can have multiple attributes. You can apply several attributes of a filter to one infoset to create a new infoset that consists of the data that you need.

### Example of filtering an infoset

A system infoset contains all files, emails, and documents of all company employees. You need to retrieve some specific data about Josh Smith: Josh's emails with a subject of `stock option` and Josh's files that are larger than 1 GB. To get these two sets of data, you can use the filter to refine the system infoset into two user infosets.

To create the first infoset about Josh's emails with the subject of `stock options`, you need to take the following steps:

- Apply a name filter attribute to find Josh Smith.
- Apply a file filter attribute to find Josh's emails with the `.MSG` extension.
- Apply an email filter attribute to find Josh's emails with a subject of `stock options`.

To create the second infoset to get Josh's files of larger than 1 GB, you need to take the following steps:

- Apply a name filter attribute to find Josh Smith.
- Apply a filter attribute of size larger than 1 GB to all of Josh Smith's files that are larger than 1 GB.

## Overlay

---

Overlays are configurable filters that display hits or matches in a selected infoset.

Within the data map, color intensifies for data objects that match the overlay change. The greater the overlay matches, the more red that tile appears within the data map.

## Set Ops

---

Set Ops allows infosets to be combined in different ways to produce another infoset.

You can select a primary infoset and use Set Ops to combine one or more infosets to create a union, intersection, symmetric difference, or subtraction infoset.

## Node ops

---

Contained data, such as data within `.ZIP`, `.TAR`, or `.PST` files, is hierarchical and can have different relationship and connections with other data. Depending on how that data is viewed, that data can give a different perception than what is represented.

The **Node ops** pane helps you understand more about what data is represented by data sets. Within Node ops, you can conduct expansion or collapse operations.

- In an expand operation, all files within an infoset are expanded, so creation of an infoset can be more accurate.

- In a collapse operation, all opened or expanded files within an infoset are collapsed, so an infoset that is created can be small.

**Note:** If the files within the infoset are not container files, then the **Expand** operator or **Collapse** operator has no effect.

## Duplicate operation

---

With Duplicate operations, you can identify varieties of duplicate data in your system. Apply filters, operations, reports, or actions to a new duplicate identification infoset to start the data deduplication process.

Duplicate operation compares objects of two infosets that are based on each other's hash value. If an object's hash value matches, the system can flag that object as a duplicate object.

## Enhancement

---

An enhancement is a way of refining or distilling an infoset. Enhancements are created as models within IBM StoredIQ Administrator. When you apply an enhancement to an infoset, it updates that infoset's index.

## Action

---

An action is an activity that is created by an administrator and conducted on an infoset. Available IBM StoredIQ Data Workbench actions include copy, copy to retention, delete, discovery export, modify attribute, move, Step-up Snippet, Step-up Full-Text, and Step-up Analytics.

Actions do not alter infosets. An infoset is a grouping of data, not the data itself, and actions are applied to the actual objects, not the infoset. When you copy, you copy the actual file, not the infoset. The same is true for copying to retention or discovery export. Actions can be scheduled to run immediately or at a predetermined time and date.

## Report

---

The reporting function provides external views of infosets and validates IBM StoredIQ processes. Reports can also be customized with the BIRT Report Designer.

You can share the information that is contained within infosets with the reporting component, which allows infosets to be transferred to other media types for review and analysis. These reports do not affect existing infosets, but provide you with more usable formats in which to understand the files and data that is captured by an infoset.

Reporting is a key step within the data-management process as it validates that processes were completed correctly within IBM StoredIQ. You can customize reports in any of these scenarios:

- Modify reports to carry your organization's custom styles and logos, aligning them with other organization-based artifacts and documentation.
- Alter the format of the content reported in existing reports. For example, you can add more columns, switch axes in a graph, or change the units for some values.
- Design reports to contain information that is not found in other, existing reports

IBM StoredIQ provides a number of preconfigured system reports, such as summaries of data objects in the system, storage use, and the number of identical documents in the system.

## Exceptions

---

When you conduct an action and encounter errors, an exception list occurs. The list helps you trace and understand what errors are so that you can correct them.

Exceptions are presented with contextual details in three areas: Events, Types, and Exception objects.

## Notices

---

This information was developed for products and services offered in the U.S.A. This material may be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing  
IBM Corporation  
North Castle Drive  
Armonk, NY 10504-1785  
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing  
Legal and Intellectual Property Law  
IBM Japan Ltd.  
19-21, Nihonbashi-Hakozakicho, Chuo-ku  
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Director of Licensing  
IBM Corporation  
North Castle Drive, MD-NC119  
Armonk, NY 10504-1785  
US

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

#### COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year).

Portions of this code are derived from IBM Corp. Sample Programs.

© Copyright IBM Corp. \_enter the year or years\_.

## Trademarks

---

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" <http://www.ibm.com/legal/copytrade.shtml>.

Adobe and PostScript are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

VMware, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Red Hat and OpenShift are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

## Terms and conditions for product documentation

---

Permissions for the use of these publications are granted subject to the following terms and conditions.

### Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

### Personal use

You may reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You may not distribute, display or make derivative work of these publications, or any portion thereof, without the express consent of IBM.

### Commercial use

You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not make derivative works of these publications, or reproduce, distribute or display these publications or any portion thereof outside your enterprise, without the express consent of IBM.

### Rights

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any information, data, software or other intellectual property contained therein.

IBM reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by IBM, the above instructions are not being properly followed.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

IBM MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

## IBM Online Privacy Statement

---

IBM Software products, including software as a service solutions, ("Software Offerings") may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering's use of cookies is set forth below.

This Software Offering does not use cookies or other technologies to collect personally identifiable information.

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek

your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, See IBM's Privacy Policy at <http://www.ibm.com/privacy> and IBM's Online Privacy Statement at <http://www.ibm.com/privacy/details> the section entitled "Cookies, Web Beacons and Other Technologies" and the "IBM Software Products and Software-as-a-Service Privacy Statement" at <http://www.ibm.com/software/info/product-privacy>.



---

# Index

## A

action [6](#), [13](#)  
AppStack [2](#)

## C

contained data [12](#)  
customized reports [13](#)

## D

data server [2](#)  
Data Server dashboard [4](#)  
Data Workbench  
    about [7](#)  
    potential uses of [7](#)  
Desktop Agent [2](#)

## E

enhancement [13](#)  
exception objects [14](#)  
exceptions [6](#), [14](#)

## F

filter [12](#)

## G

gateway [2](#)

## H

harvest [11](#)

## I

IBM StoredIQ Administrator [5](#)  
IBM StoredIQ Data Server [4](#)  
IBM StoredIQ Data Workbench [6](#)  
IBM StoredIQ Desktop Data Collector [9](#)  
IBM StoredIQ index [11](#)  
index types  
    full-text index [11](#)  
    metadata index [11](#)  
information set [11](#)  
infoset  
    system infoset [11](#)  
    user infoset [11](#)  
intersection [12](#)

## K

key terms  
    action [11](#)  
    filter [11](#)  
    index [11](#)  
    information set [11](#)  
    infoset [11](#)  
    overlay [11](#)  
    report [11](#)  
    scope operation [11](#)  
    set operation [11](#)

## L

legal  
    notices [15](#)

## N

Node ops  
    Collapse [12](#)  
    Expand [12](#)  
notices  
    legal [15](#)

## O

overlay [12](#)

## R

report  
    report types [13](#)  
reports [13](#)

## S

Set Ops [12](#)  
subtraction [12](#)  
symmetric difference [12](#)

## U

union [12](#)

