Redbooks
ibm.com/redbooks

IBM StoredIQ Introduction and Planning Considerations

Terry Ellis

Analytics

ECM

IBM

Redpaper

IBM StoredIQ Introduction and Planning Considerations

The IBM® StoredIQ® Platform 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 e-discovery, records retention and disposition, compliance, and storage optimization initiatives.

This IBM Redpaper™ publication provides an introduction to the StoredIQ Platform and outlines basic planning considerations.

Introduction to IBM StoredIQ

IBM StoredIQ provides valuable insight into unstructured content across many data sources within the organization.

This introduction describes the following topics:

- Business issues and challenges
- Product overview
- Use cases

Business issues and challenges

Unstructured data is growing at an unprecedented rate within most organizations. This section describes some of the common business challenges faced by most customers.

Unstructured data growth

Discussions in the industry have revolved around information under management doubling every two years. This content is typically spread across many disparate data sources, and IT, legal, and records management have little knowledge of what purpose the content servers or what regulations govern its lifecycle.

Disparate data sources

Most enterprises house unstructured data in a variety of locations: file shares, Microsoft SharePoint, mail servers, mail archives, ECM repositories, and cloud-based repositories. Having various locations presents several challenges. Applying consistent policy across the data sources is difficult. Searching these repositories for content subject to legal review or record retention policies becomes burdensome, costly, and often is not done.

When repositories are searched for content, it must be done with a variety of tools native to each repository. This leads to inconsistent results across data sets, becomes burdensome to the organization, and requires involvement by many different data experts.

Cost

The industry has also discussed value estimates, which show that over half of the retained unstructured data in an organization has no value and there is no regulatory or legal reason to retain it. This over-retention of data can become costly for an organization. Although a disk might be inexpensive, the cost associated with acquisition, provisioning, maintenance, upgrades, and administration can consume the majority of an IT budget. Projecting these costs out several years shows an alarming trend. The cost to maintain storage at its current growth rate completely consumes or outpaces projected future budget. This cost severely limits the ability for organizations to take on new initiatives to grow and enhance the business.

Beyond the cost of storage, significant costs are associated with searching and culling of extraneous data. When legal matters arise, content must be searched, culled, collected, and reviewed. Over-retention of data extends the level of effort required to identify relevant content and typically results in over-production for review. Collected data must be reviewed; this is a manual, labor-intensive step. Estimates in the industry have shown that review costs are in the tens of thousands of US dollars per gigabyte. Proactively dispositioning data according to business, record, and legal policies is the most effective way of containing this cost.

Risk

Sensitive personal information (SPI), like customer account information, in secondary ungoverned data sources puts companies at risk for proliferation, co-mingling, disclosure, inadvertent access, or other mishandling with significant financial, operational, and reputational impact. Data breaches can occur from insider theft of SPI that was not properly identified and secured.

During a legal matter, data, if still in possession and demanded by an adverse party in e-discovery, must be produced. This reduces agility to respond timely to governmental requests. Previously collected data of a sensitive or harmful nature that is not dispositioned upon matter closure is at continued risk of disclosure.

When data retention and disposal are exercised in an ad hoc fashion by employees and lines of businesses against ungoverned data, this uneven enforcement of records policy puts defensibility of entire records and compliance programs at risk. Companies might face fines for failure to establish and enforce compliance.

Product overview

This section provides a brief overview of the IBM StoredIQ Platform, interfaces, terminology, process, and use cases.

IBM StoredIQ interface components

The IBM StoredIQ interface components include IBM StoredIQ Platform Data Server, IBM StoredIQ Administrator, IBM StoredIQ Data Workbench, IBM StoredIQ Data Script, IBM StoredIQ Policy Manager, and IBM StoredIQ Desktop Data Collector.

IBM StoredIQ Platform Data Server

IBM StoredIQ Platform 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 monitors and manages the distributed infrastructure at a client site. IBM StoredIQ Administrator sits between the IBM StoredIQ Platform interface and the applications and facilitates the transfer and communication of information. IBM StoredIQ Administrator understands and manages IBM StoredIQ Platform concepts such as volumes, indexes, harvests, and configurations. At the same time, it manages the application concerns such as infoset lifecycle and creation, volume configuration, and action and target set management. To this end, it is divided into two sections (platform and application) so that the administrators know where to accomplish a task.

IBM StoredIQ Data Workbench

With IBM StoredIQ Data Workbench, you can visualize the indexed data and identify potential "red-flag" issues to know how much and what types of data you have on different types of servers. It alerts people about potentially interesting or useful data. It helps ensure that the data of an enterprise is an asset, not a liability.

IBM StoredIQ Data Script

IBM StoredIQ Data Script automates execution within IBM StoredIQ Platform. Therefore, you can script, automate, and monitor processes that otherwise normally are manual processes that are run within IBM StoredIQ Data Workbench. IBM StoredIQ Data Script focuses on repeatable, understood, and approved processes for the purposes of culling and refining data in an approved manner.

IBM StoredIQ Policy Manager

IBM StoredIQ Policy Manager acts on data in an automatic fashion at scale, running policies that affect data objects without requiring review.

IBM StoredIQ Desktop Data Collector

IBM StoredIQ Desktop Data Collector deploys from the IBM StoredIQ Platform Data Server Administrator interface. It indexes desktops as volumes. The volumes appear in the Data Server Administrator interface and IBM StoredIQ Data Workbench, where the data can be analyzed and acted upon.

IBM StoredIQ Platform uses a non-invasive deployment, specifically these items:

- Lightweight client deployment with no browser plug-ins and no client-side UI installations. All major browsers, such as Internet Explorer and Firefox, are supported natively.
- One (common) index that is shared across all use cases. This index supports a wide variety of data sources and is used by application dashboards for many use cases.
- No agents are placed on data sources. The native API or protocol is used whenever possible. Third-party or custom connectors are used only when necessary.

> **Note:** Agents are used for desktop collection.

Terminology

This section explains several frequently used terms in StoredIQ.

Volume

A volume represents a data source or destination that is available in the network to the IBM StoredIQ Platform appliance. A volume can be a disk partition or group of partitions that is available to network users as a single designated drive or mount point. IBM StoredIQ Platform volumes have the same function as partitions on a hard disk drive. When you format the hard disk drive on your PC into drive partitions A, B, and C, you are creating three partitions that function like three separate physical drives. Volumes behave the same way that hard disk partitions behave. You can set up three separate volumes that originate from the same server or across many servers. Only administrators can define, configure, and add or remove volumes to IBM StoredIQ Platform.

Primary volume

Primary volumes can be created as data sources using these volume types: CIFS, NFS v2 and v3, Exchange, SharePoint, Documentum, Discovery Accelerator, IBM Domino®, IBM FileNet®, NewsGator, Livelink, Jive, Chatter, IBM Content Manager, and CMIS.

Retention volume

Retention volumes store data objects that are placed under retention, which means that the object is retained. Retention volumes can be added and configured. Applicable volume types include Enterprise Vault, CIFS (Windows platforms), NFS v3, Centera, and Hitachi.

System volume

System volumes support volume export and import. When you export a volume, data is stored on the system volume. When you import a volume, data is imported from the system volume.

Discovery export volume

Discovery export volumes contain the data produced from a policy, which is kept so that it can be exported as a load file and uploaded into a legal review tool. Administrators can also configure discovery export volumes for managing harvest results from cycles of a discovery export policy.

Indexes

When you define volumes, you can determine the type and depth of index that is conducted.

Three levels of analysis are as follows:

- *System metadata index*: This level of analysis runs with each data collection cycle and provides only system metadata for system data objects in its results. It is useful as a simple inventory of what data objects are present in the volumes you defined and for monitoring resource constraints (such as file size) or prohibited file types (such as `.MP3`).
- *System metadata plus containers*: In a simple system metadata index, items within container data objects (compressed files, PSTs, emails with attachments, and the like) are not included. This level of analysis provides container-level metadata in addition to the system metadata for system data objects. All objects within containers are included in the index.
- *Full-text and content tagging*: This option provides the full local language analysis that yields the more sophisticated entity tags. Naturally, completing a full-text index requires more system resources than a metadata index. Users must carefully design their volume structure and harvests so that the maximum benefit of IBM StoredIQ Platform's sophisticated analytics are used, but not on resources that do not require them. Parameters and limitations on full-text indexing are set when the system is configured.

Harvest

A harvest is a job that indexes a volume. Harvests can be run immediately or scheduled.

Object

Object refers to a single item indexed by the system. This can be a document, container, object within a container, media file, and so on.

Information set (Infoset)

An infoset is a collection of volumes and the indexes that are created from them. Various types of infosets exist:

- All Data Objects infoset

  The All Data Objects infoset contains index information about all data objects that are harvested. This includes any container objects harvested. The total object count and size represent the expanded view of all objects as though they were on disk.

- All System Level Objects infoset

  The All System Level Objects infoset contains all objects indexed, but does not include content within container files. The object count and size represent the actual size on disk.

- System infoset

  A System infoset can be created in IBM StoredIQ Data Workbench to allow users to have a different starting point than the All Data Objects infoset. System infosets are created in the IBM StoredIQ Administrator interface and represent a group of chosen volumes. System infosets allow users to group together volumes that might represent things like All CIFS, All HR Files (made up of HR department CIFS, HR SharePoint sites, and so on.)

- User infoset

  The User infoset is created in the IBM StoredIQ Data Workbench interface by applying a filter to an existing System or User infoset. The resulting set represents the set of objects that meet the filter criteria.

Process overview

This section provides a basic overview of the process of analyzing and acting on unstructured data. It outlines the basics of each of the major sections within the IBM StoredIQ Data Workbench user interface.

Data source indexing

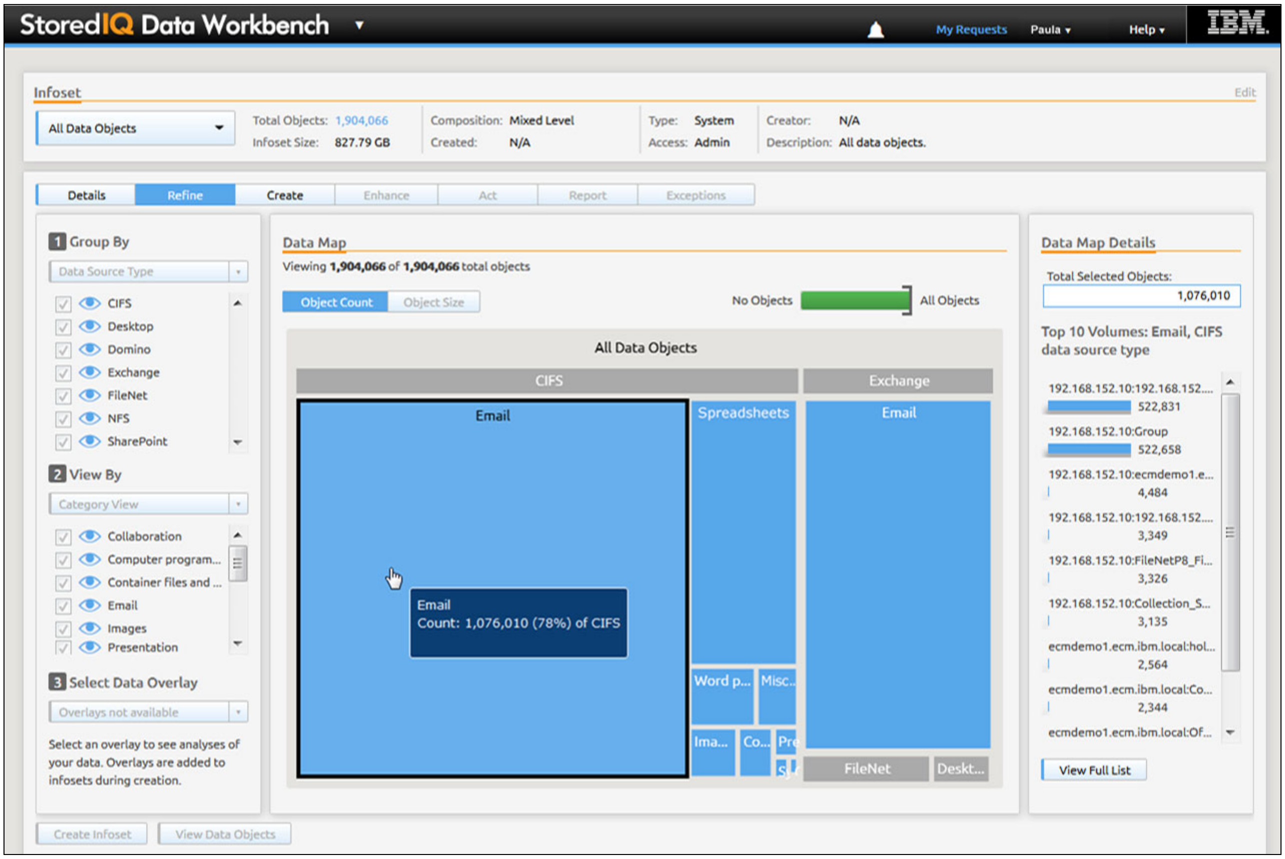
Analyzing data with IBM StoredIQ begins with the indexing of the data source. A primary volume is created using the IBM StoredIQ Administrator or IBM StoredIQ Platform Data server interface. This volume maps to a data source such as a CIFS, NFS, or SharePoint site. A harvest job is created and runs, indexing content and adding information about the each data object to the IBM StoredIQ data base. When indexing is complete, this information is available for analysis in the IBM StoredIQ Data Workbench.

Infoset details

The infoset Details tab in the IBM StoredIQ Administrator shows the number of objects and total size of the chosen information set. It also shows the number of top level objects and child objects if container level indexing was performed. A list of the data objects within the infoset can also be displayed.

Refine

The Refine tab (Figure 1) displays a topographical layout of the selected infoset. The interface displays the list of data sources; the user can then analyze the infoset based on category, created date, last modified date, or size.



Infoset

All Data Objects

Total Objects: 1,904,066
Infoset Size: 827.79 GB
Composition: Mixed Level
Created: N/A
Type: System
Access: Admin
Creator: N/A
Description: All data objects.

Data Map

Viewing 1,904,066 of 1,904,066 total objects

Email
Count: 1,076,010 (78%) of CIFS

Data Map Details

Total Selected Objects: 1,076,010

Top 10 Volumes: Email, CIFS data source type

192.168.152.10:192.168.152.... 522,831
192.168.152.10:Group 522,658
192.168.152.10:ecmdemo1.e... 4,484
192.168.152.10:192.168.152.... 3,349
192.168.152.10:FileNetP8_Fi... 3,326
192.168.152.10:Collection_S... 3,135
ecmdemo1.ecm.ibm.local:hol... 2,564
ecmdemo1.ecm.ibm.local:Co... 2,344
ecmdemo1.ecm.ibm.local:Of...

Select an overlay to see analyses of your data. Overlays are added to infosets during creation.

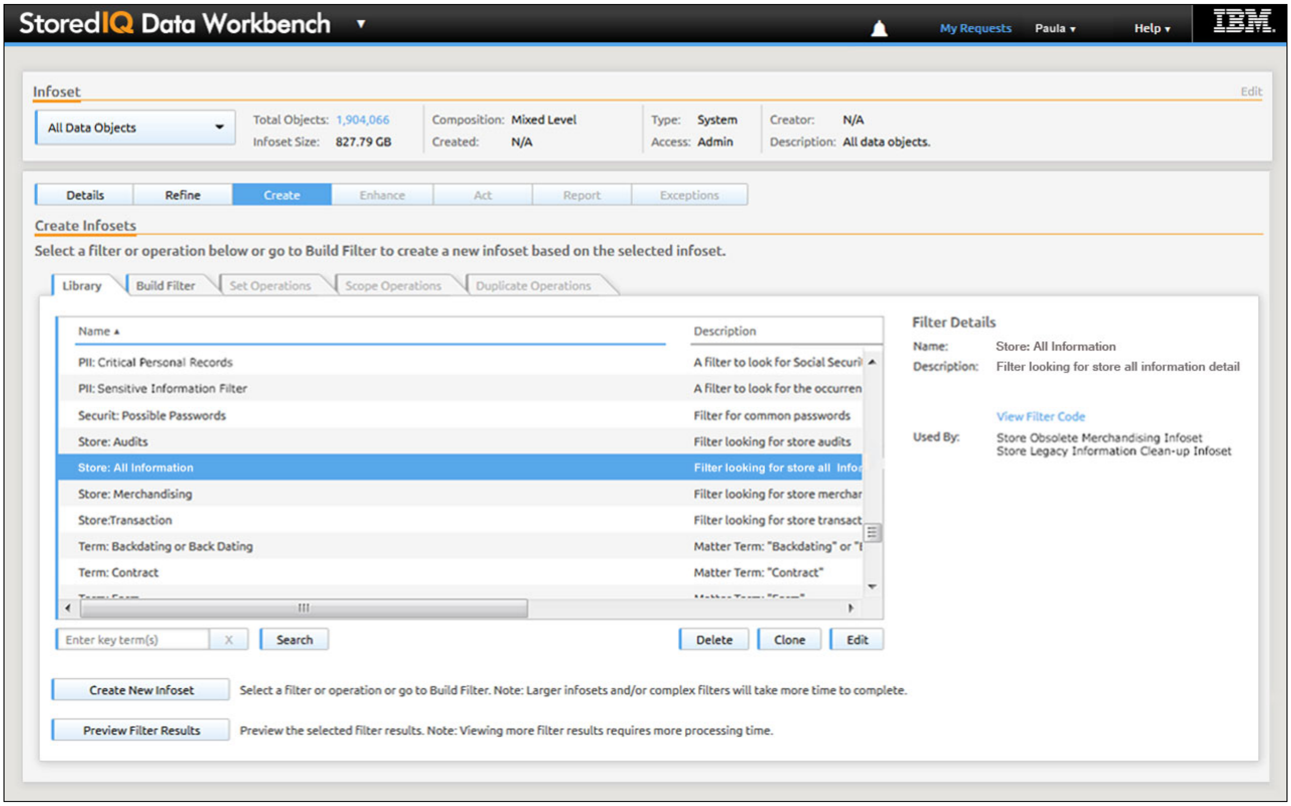
Figure 1 StoredIQ Data Workbench Refine tab

Information is displayed in the Data Map panel as a set of boxes of various sizes. The size of the box represents the relative percentage of what the box represents.

Data Map Details panel displays the count of objects by individual volume for the top 10 volumes.

Create

The Create tab (Figure 2 on page 7) is where users create filters or choose a filter previously created to apply to the infoset. Figure 2 on page 7 shows a list of previously created filters.



Infoset

All Data Objects

Total Objects: 1,904,066
Infoset Size: 827.79 GB
Composition: Mixed Level
Created: N/A
Type: System
Access: Admin
Creator: N/A
Description: All data objects.

Create Infosets

Select a filter or operation below or go to Build Filter to create a new infoset based on the selected infoset.

Name ▲	Description
PII: Critical Personal Records	A Filter to look for Social Securi
PII: Sensitive Information Filter	A Filter to look for the occurren
Securit: Possible Passwords	Filter for common passwords
Store: Audits	Filter looking for store audits
Store: All Information	Filter looking for store all Info
Store: Merchandising	Filter looking for store merchar
Store:Transaction	Filter looking for store transact
Term: Backdating or Back Dating	Matter Term: "Backdating" or "I
Term: Contract	Matter Term: "Contract"

Filter Details
Name: Store: All Information
Description: Filter looking for store all information detail
Used By: Store Obsolete Merchandising Infoset
Store Legacy Information Clean-up Infoset

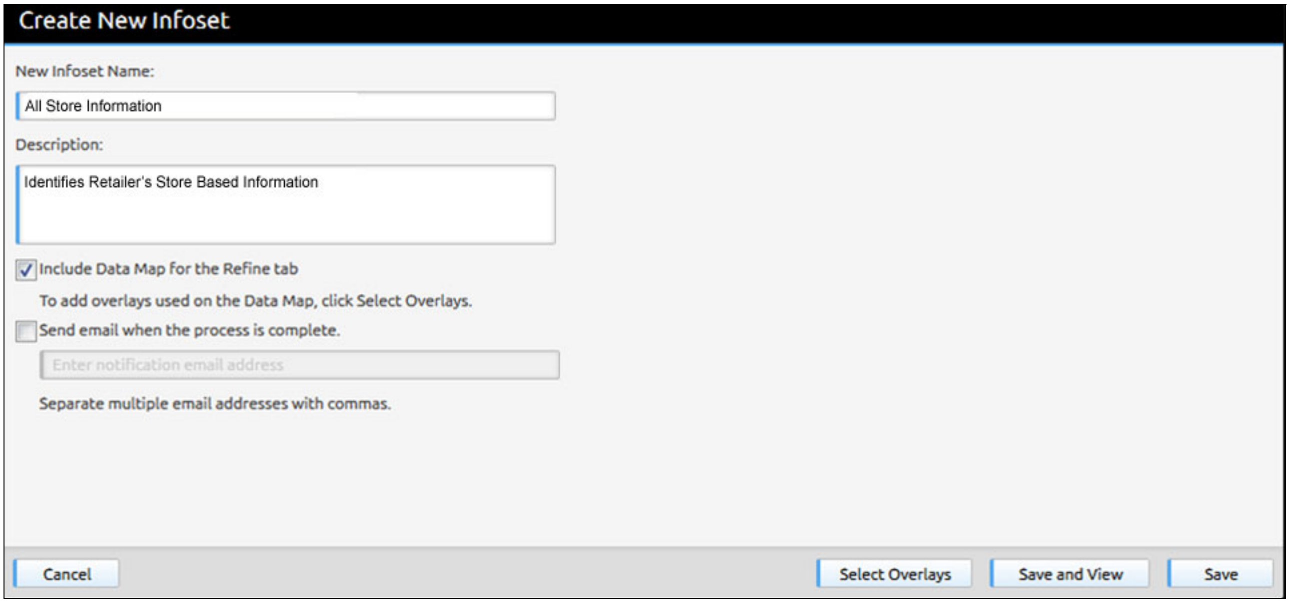
Enter key term(s)

Create New Infoset: Select a filter or operation or go to Build Filter. Note: Larger infosets and/or complex filters will take more time to complete.

Preview Filter Results: Preview the selected filter results. Note: Viewing more filter results requires more processing time.

Figure 2 StoredIQ Data Workbench Create tab, filter Library

A new infoset is created by applying the filter to the existing infoset. This process produces a new infoset containing information about only the objects that meet the filter criteria. (Figure 3).



New Infoset Name:
All Store Information

Description:
Identifies Retailer's Store Based Information

☑ Include Data Map for the Refine tab

To add overlays used on the Data Map, click Select Overlays.

☐ Send email when the process is complete.

Enter notification email address

Separate multiple email addresses with commas.

Figure 3 Creating a new infoset

Enhance

If data must be indexed for classification purposes, it must be first enhanced by running the Step-Up Snippet. This step calculates relevancy rankings for each document compared to the chosen classification model.

Act

With the Act tab, a user can take action on the contents of the information set. Actions are created in the IBM StoredIQ Administrator interface and run here in the IBM StoredIQ Data Workbench. Actions can be created to move, copy, delete, export, or retain copy data. The action can either be run immediately or scheduled for a later time.

If data was originally indexed with the metadata-only option, it can be full-texted, indexed by running the Full Text Step Up. After complete, the infoset can be further analyzed and refined based on the content of the objects.

Report

The Report tab contains a list of all of the ready-for-use reports that are included in IBM StoredIQ. Running a report on an infoset can provide details to the user about information such as data topology, occurrence of key terms, and summary of duplicate items. The list of objects along with their metadata can also be exported from the system into a `.csv` file.

Exceptions

The Exceptions tab lists any exceptions that might occur during processing. These errors can include, for example, file copy or delete errors that might have occurred during the action phase.

Use cases

IBM StoredIQ Platform can address many use cases within an organization. This section lists the most common. One of the primary benefits of IBM StoredIQ is that all use cases use the same index and can be addressed simultaneously, providing value to many stakeholders within the organization.

Data assessment

Assessment starts with understanding your data. By indexing source data, IBM StoredIQ helps you to begin to understand your data footprint. When you begin to understand the topology of the data, you can begin to filter it for use cases described here.

Initially, a metadata index of content can be performed. This level of indexing captures file properties such as name, path, size, created, accessed and last modified dates, and so on. Information is presented in the data map on the Refine tab. This topology map shows how the data is distributed across the system by category, size, and age range. Here, you can quickly begin to spot potential policy violations or opportunities for storage reduction.

In addition, data (typically a subset) can be full-text indexed. There is usually no value in full-text indexing many categories of content such as computer files, system files, media files, and files over a certain age. By eliminating these types of files from the full text index, processing efficiencies, time, and disk space are saved.

During data assessment, filters can be created to identify data that might be outside of corporate policies, as in these examples:

- Data that is beyond corporate retention policies
- Data outside of acceptable use policies such as multimedia files
- Data owned by employees no longer with the company
- Data containing personally identifiable information (PII), personal credit information (PCI), and personal health information (PHI)
- Duplicate content

In the data assessment use case, data is only indexed, analyzed, and reported on but no action is taken.

Data cleanup

After data is assessed, data cleanup entails finding relevant sets of data that can be deleted from the data source. As described in the data assessment use case, filters are created to identify content that no longer has business value. After data is identified, IBM StoredIQ can be used to remove the content by employing a delete action.

Often companies choose to identify content they feel is eligible for depletion, distribute detail reports to lines of business, and then wait a period of time before actual deletion occurs. In this case, data can be kept in place for the desired period of time and the modified date checked before actual deletion or, in some cases, data is relocated to a staging or quarantine area while waiting to be deleted.

Data cleanup generally occurs in the following stages:

- Identify and remove trivial content. Trivial content is content that has no business value and might include computer files, system files, multi-media files, and other personal content.
- Identify and remove obsolete content. Much of the content that exists within an organization is retained beyond its usefulness. Departmental retention policies can be applied and data that is kept beyond the longest retention policy can quickly be identified. Often this content can be removed from the system without in-depth records classification or application of records policies.
- Identify and address duplicate content. Duplicate data reports can be produced showing all exact duplicates within the data source. Although duplicate content certainly exists, the report helps identify how large of a problem it is and helps users prioritize how and when to address the issue.

Compliance remediation

Sensitive data that might contain information such as social security numbers, credit card numbers, proprietary information, customer data, and so on can pose great risk to an organization if it is not properly secured. IBM StoredIQ can both identify this content also help remediate it. IBM StoredIQ includes macros to identify US Social Security numbers (SSNs), credit cards, phone numbers, and several other types of PII and PCI. Filters can be created to identify other types of sensitive information within the organization such as account numbers, policy number, customer numbers, and so on.

These filters can be applied to data sources where sensitive information should not exist. This then provides identification of compliance policies. After data is identified, reports can be produced to remediate users and correct behavior. Data can also be relocated to the appropriate data store or deleted.

E-discovery

One of the primary use cases for IBM StoredIQ is the identification of content subject to a legal discovery. Data experts can find relevant data located in several disparate repositories with the use a single tool. Searches can be configured to find content based on dates, owners, key words, word proximity, and so on. After data is identified, it can be presented to the legal team before collection. This allows the legal team to perform an early data assessment and better understand both the risk and the burden associated with the case. The infoset can also undergo additional refinement to further cull it to a more accurate set of data that might need to be reviewed.

If collection is necessary, IBM StoredIQ can be used to copy data from its original source to a hold location or export the content into one of several industry standard formats for further review.

Records identification

Corporations have often relied on users to understand corporate retention schedules and to properly identify corporate records. This places a great burden on the user and typically has inconsistent results. By using StoredIQ to identify data outside the records repository, this process can be greatly improved. Data can be filtered, culled, and classified according to metadata, key words, and also classified by using the integration with IBM Content Classification. This way allows for more intelligent classification of content by the use of natural language processing. This way also allows StoredIQ to identify corporate records based on the content of the document and applies the classification in a standard process. After records are properly identify, they can be relocated to a records repository, if you want, and also dispositioned when the retention period is reached.

Planning considerations

The planning process for the deployment of IBM StoredIQ is described in the following topics:

- IBM StoredIQ components
- Open Virtual Architecture (OVA) configuration requirements
- Network and port requirements
- Environment sizing guidelines
- Deployment architecture

IBM StoredIQ components

The three components of the IBM StoredIQ solution are the application stack, the gateway, and the data server. These components work together as the IBM StoredIQ products.

Application stack

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

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

IBM StoredIQ Platform Data Server helps you to understand the data landscape of the enterprise. It 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, and file owners. The data servers push 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. In addition to an administrator user interface, administrators can deploy the IBM StoredIQ Desktop Data Collector and index desktops from the data server.

IBM StoredIQ Platform uses a non-invasive deployment, specifically these items:

- Lightweight client deployment with no browser plug-ins and no client-side UI installations. All major browsers, such as Internet Explorer and Firefox, are supported natively.
- One (common) index that is shared across all use cases. This index supports a wide variety of data sources and is used by application dashboards for many use cases.
- No agents are placed on data sources. The native API or protocol is used whenever possible. Third-party or custom connectors are used only when necessary.

Note: Agents are used for desktop collection.

Open Virtual Architecture (OVA) configuration requirements

IBM StoredIQ Platform is deployed as virtual appliances and currently only supported in VMware ESXi 5.x or later environments. You must have a virtual infrastructure that meets the IBM StoredIQ Platform hardware requirements.

Application stack

The application stack has these hardware requirements:

- vCPU: 1
- Memory: 4 GB
- Storage:
  - Primary disk (vmdisk1): 21 GB
  - Data disk (vmdisk2): 10 GB

Gateway server

The gateway server has these hardware requirements:

- vCPU: 2
- Memory: 8 GB
- Storage:
  - Primary disk (vmdisk1): 100 GB
  - Data disk (vmdisk2): 75 GB
  - Swap disk (vmdisk3): 40 GB

Data server

The data server has these hardware requirements:

- CPU: 4

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

- Memory: 16 GB

  Although the minimum value works under a light-load condition, as the load increases, the data server quickly starts consuming swap space. For high-load situations, increasing memory 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 virtual machine disk (VMDK) that contains the IBM StoredIQ operating code. Do not change its size.

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

  – Data disk (vmdisk2, SCSI 0:1): Default is 1.9 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. The default data disk size is therefore 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 many 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 deliver acceptable performance. In the situations when the load on the system is high, the IOPS that is used can reach up to 7000 with main write operations.

Network and port requirements

The number and types of data sources can drastically impact the scale and scope of what needs to be deployed. The complexity of the source directly affects the number of data servers to be deployed, for example, exchange versus simple text documents in a CIFS location.

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 ports 8765 and 5432
- Ports 80, 443, and 22 from the administrative user's workstation (place from which the administrator is completing work with IBM StoredIQ Administrator) to the application stack and data server IP addresses
- Port 22 from the administrative workstation to the gateway IP address

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:

- One data server per 30 TB of file shares. This varies depending on number of volumes, objects per volume, and object types.
- One gateway per 50 data servers.
- One application server.
- NFS is slightly faster than CIFS for metadata only, but assume timings are equal for this sizing discussion.
- Full-content processing of file (for example, `.ZIP`, `.RAR`, `.GZ`) and email archive (`.PST`, `.NSF`, `.EMX`) processing are slower because items must be extracted from the archives. If a significant number of these files are 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.
- A bytes/time metric is appropriate for metadata-only processing that computes a hash and full-content. 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.
- For a metadata-only computing a hash processing, 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 an 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.

- A 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 an 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” can be set to eight for optimal performance.
- The interrogator process count for all other processing that involves reading all content is assumed to be four per data server.
- The interrogator count can be viewed as the number of client connections that are made to a data source that is 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 a local area network). The data servers need a low-latency high-bandwidth connection to an 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 or more are acceptable.

VMware vSphere requirements

The VMware vSphere used in a StoredIQ environment should meet the following requirements:

- VMware vSphere V5.x or later
- VMware virtual machine version 8.0 or later
- VMware license to enable the required processor cores and memory for the virtual machine

Deployment architecture

The IBM StoredIQ Platform is deployed as a set of virtual appliances, which greatly enhances scalability and flexibility, and reduces the time to value.

Architecture example

Figure 4 shows a sample architecture. It depicts a single application server, a single gateway server, and multiple data servers. This typical architecture allows indexing of multiple data sources across the enterprise.

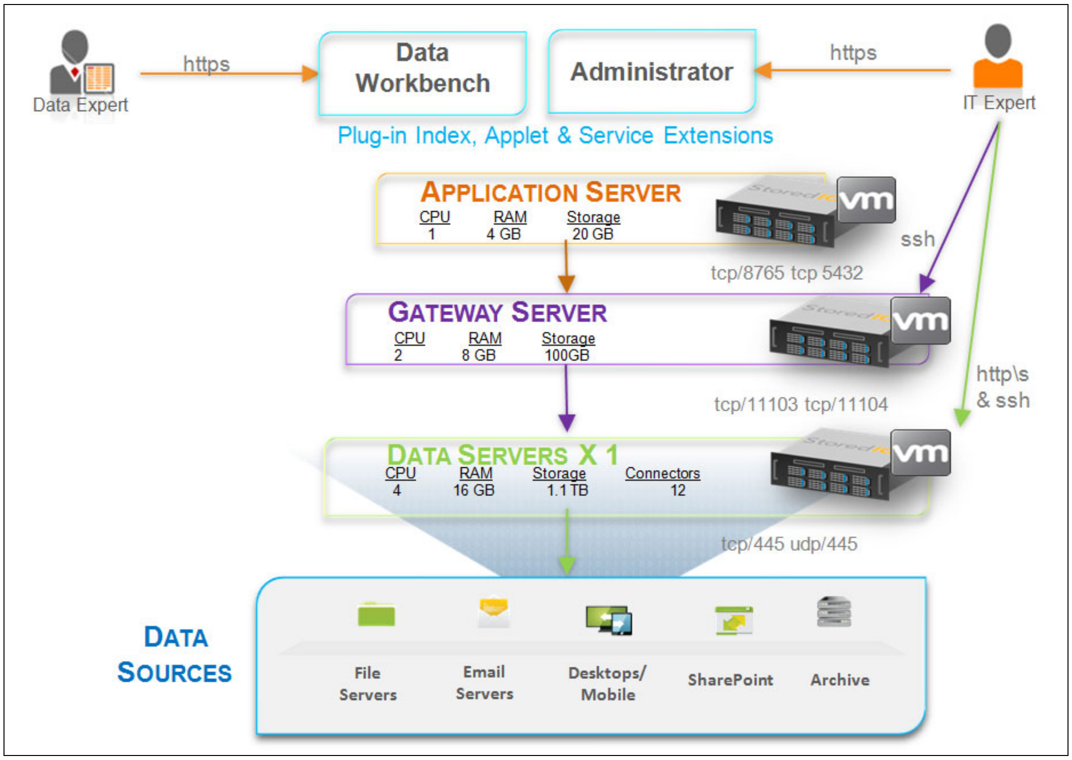


Figure 4 Sample architecture

Data server distribution considerations

Consider the following information when you are determining how to deploy the number and location of the IBM StoredIQ data servers:

- Volume of data
- Distribution of data
- Speed of indexing that you want

Volume of data

A data server is required for approximately 150 million objects, which estimates to be approximately 30 TB. As the data volume approaches this threshold, add more data servers.

> **Note:** The amount of data that a data server can manage is actually determined by the number of objects. If you are processing container files, the number of objects managed can be orders of magnitude higher than the actual number of objects on the data source. For example, a single PST file might contain hundreds of thousands of emails, attachments, and so on.

Distribution of data

For purposes of indexing efficiency and reduction of network traffic, a good approach is to have data servers geographically co-located with the data source they are indexing. If deploying IBM StoredIQ against data in multiple data centers, place data servers in each data center also. Multiple data servers per geographical location is necessary for managing more than 30 TB of data per data center.

When small amounts of data are distributed across a number of geographical locations, such as regional offices, placing data servers in each location might be impractical. The data server can index and manage data from a remote location; however, indexing times will lengthen significantly.

Speed of indexing that you want

Indexing speed on a data server can be limited by the number of interrogator processors, memory, disk speed, and access to the source data. Configuring multiple data servers to index content in parallel can greatly increase the data that can be indexed in a given time. Careful planning must occur to ensure no overlap in targeted data exists among multiple data servers.

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Terry Ellis is an Information Lifecycle Governance Technical Subject Matter Expert in the United States. Terry has been working in the Information Lifecycle Governance (ILG) area for 12 years. He worked for three years in Lab Services delivering ILG products. He then worked as a worldwide ECM Architect before joining the StoredIQ team three years ago. Terry currently works with customers who are interested in maturing their ILG practices.

Thanks to the following people for their contributions to this project:

Jim Sikes, WW ILG Tech Sales Leader, IBM System
Whei-Jen Chen, Project Leader, IBM International Technical Support Organization

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Learn more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new IBM Redbooks® publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the U.S.A.

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.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: 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 states 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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites 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.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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.

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.

This document REDP-5315-00 was created or updated on December 22, 2015.

Send us your comments in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:
  ibm.com/redbooks
- Send your comments in an email to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Domino®
FileNet®
IBM®
Redbooks®
Redpaper™
Redbooks (logo) ®
StoredIQ®

The following terms are trademarks of other companies:

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

Other company, product, or service names may be trademarks or service marks of others.

REDP-5315-00

ISBN 738454850

Printed in U.S.A.