

IBM Spectrum Scale and IBM StoredIQ

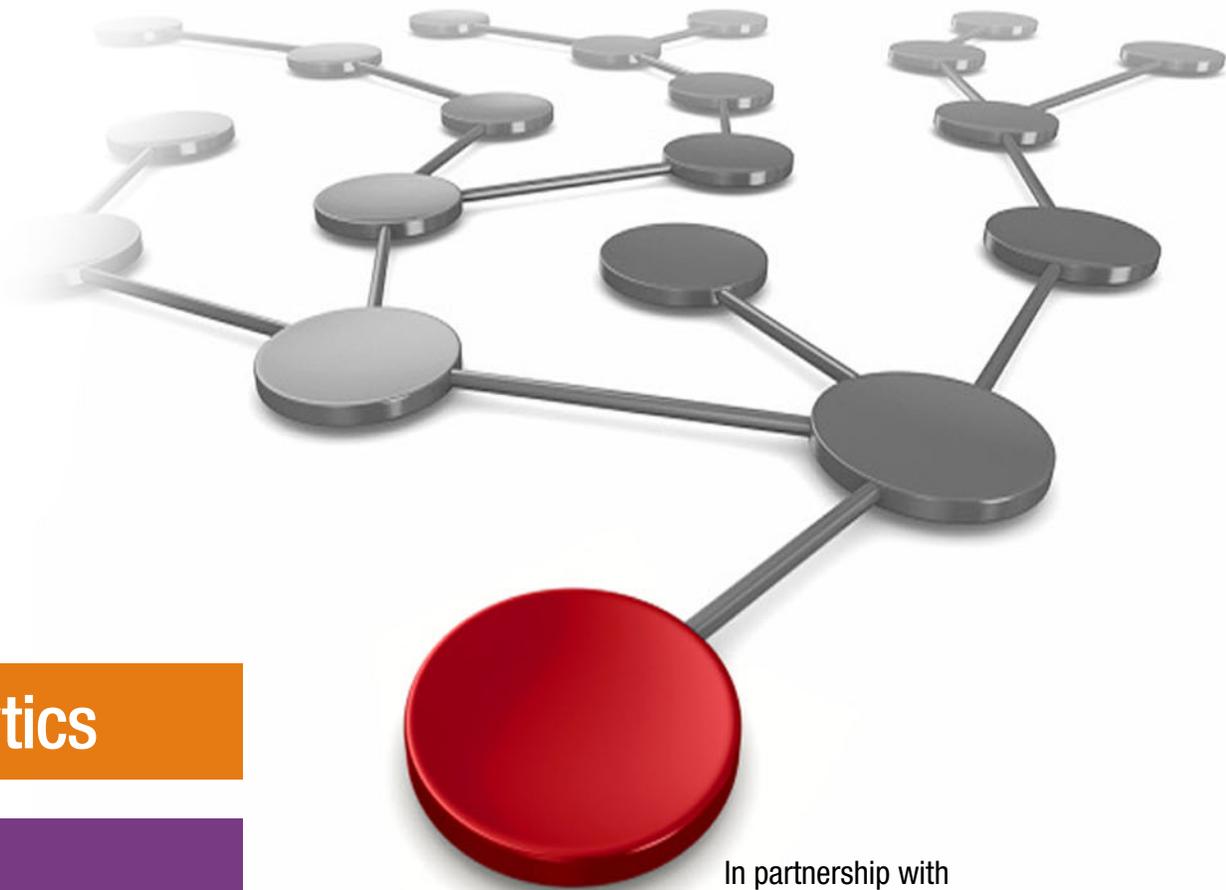
Identifying and securing your business data to support regulatory requirements

Sandeep R Patil

Sasikanth Eda

Atul V Gore

Nils Haustein



 Analytics

Storage

In partnership with
IBM Academy of Technology



IBM Spectrum Scale and IBM StoredIQ Overview

Having the appropriate storage for hosting business critical data and the proper analytic software for deep inspection of that data is becoming necessary to get deeper insights into the data so that users can categorize which data qualifies for compliance.

This IBM® Redpaper™ publication explains why the storage features of IBM Spectrum™ Scale, when combined with the data analysis and categorization features of IBM StoredIQ®, provide an excellent platform for hosting unstructured business data that is subject to regulatory compliance guidelines, such as General Data Protection Regulation (GDPR).

In this paper, we describe how IBM StoredIQ can be used to identify files that are stored in an IBM Spectrum Scale™ file system that include personal information, such as phone numbers. These files can be secured in another file system partition by encrypting those files by using IBM Spectrum Scale functions. Encrypting files prevents unauthorized access to those files because only users that can access the encryption key can decrypt those files.

This paper is intended for chief technology officers, solution, and security architects and systems administrators.

Introduction to IBM Spectrum Scale

IBM Spectrum Scale is a proven, scalable, high-performance file system that is suitable for various use cases. It provides world-class storage management with extreme scalability, flash accelerated performance, and automatic storage tiering capabilities. IBM Spectrum Scale reduces storage costs while improving security and management efficiency in cloud, big data, and analytics environments. IBM Spectrum Scale provides the following benefits:

- ▶ Virtually limitless scaling to nine quintillion files and yottabytes of data.
- ▶ High performance and simultaneous access to a common set of shared data.
- ▶ Integrated information lifecycle management (ILM) functions to automatically move data between storage tiers including flash, disk, tape, and Object Storage (public and private cloud). This benefit can dramatically reduce operational costs as fewer administrators can manage larger storage infrastructures.
- ▶ Software-defined storage allows you to build your infrastructure solution with the following characteristics:
 - Easy to scale with relatively inexpensive commodity hardware while maintaining world-class storage management capabilities.
 - Deployable on Amazon Cloud (AWS) and IBM Cloud™.

- Cross-platform solution available on IBM AIX®, Linux, and Windows server nodes, or a mix of all three. IBM Spectrum Scale is also available for IBM Z®.
- ▶ Available as the pre-packaged storage solution IBM Elastic Storage™ Server with declustered RAID included.
- ▶ Global data access across geographic distances and unreliable WAN connections.
- ▶ Multi-site support, which connects a local IBM Spectrum Scale cluster to remote clusters to provide greater administrative flexibility and control.
- ▶ Proven reliability, even across multiple sites and support for concurrent hardware and software upgrades.
- ▶ State-of-the-art protocol access methods for managing files and objects under the same global namespace. This feature makes more efficient use of storage space and avoids data islands. The supported protocols include NFS, SMB, POSIX, OpenStack Swift, and S3.
- ▶ Seamless integration for Hadoop applications by way of the HDFS Transparency feature.
- ▶ Proven security features to ensure data privacy, authenticity, and auditability.
- ▶ File level encryption for data at rest and secure erase.
- ▶ Policy-driven compression to reduce the size of data at rest and increase storage efficiency.
- ▶ Supports OpenStack deployments through its cinder, manila, and glance driver support.
- ▶ Can be used as permanent storage for Docker Containers.
- ▶ Includes a GUI to simplify storage administration tasks and monitor many aspects of the system.

An overview of IBM Spectrum Scale is shown in Figure 1.

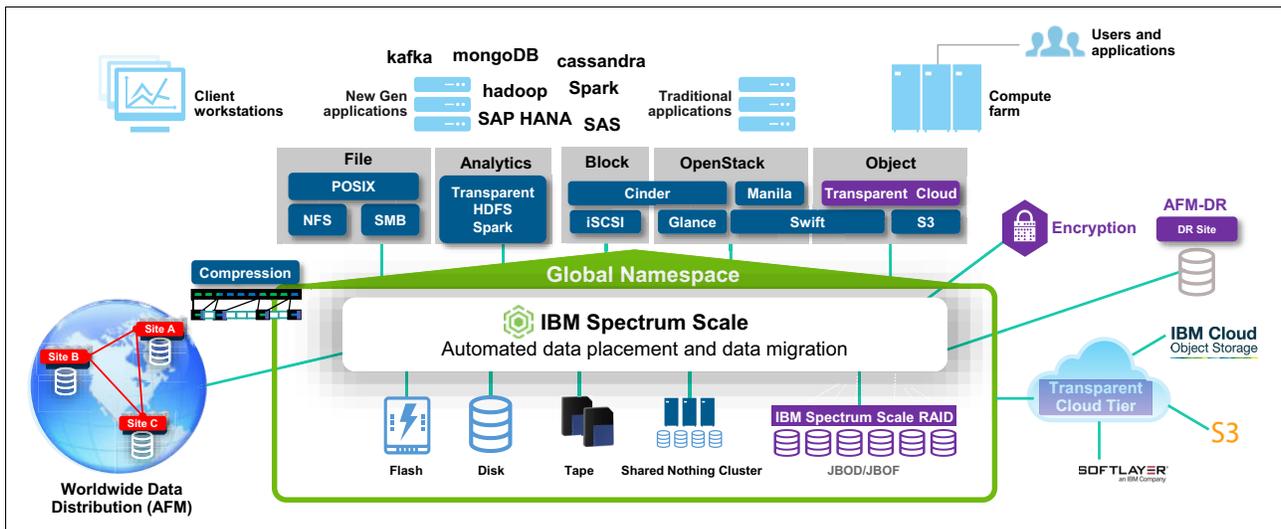


Figure 1 IBM Spectrum Scale overview

IBM Spectrum Scale is especially used in high-performance and computationally demanding environments across different branches including banking, financial, healthcare, oil and gas, and automotive industries.

Note: IBM Spectrum Scale is available in a no-cost trial version that runs in a virtual environment. The trial version includes a fully preconfigured IBM Spectrum Scale instance in a virtual machine that is based on IBM Spectrum Scale version 5.0. For more information, see this web page:

<https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage>

Benefits of IBM Spectrum Scale Storage for GDPR

The European Union (EU) General Data Protection Regulation (GDPR) focuses on the protection of personal data (article 4, section 1) and must be met by any organization that collects and stores personal data from individuals who are in any of the 28 EU member states. Because personal data that is subject to GDPR is also stored in an unstructured data format, a scale-out file system, such as IBM Spectrum Scale, provides the following essential functions to meet GDPR requirements:

- ▶ Provides a single global namespace that can store, manage, and protect unstructured data.
- ▶ Offers various protocol access methods, including NFS, SMB, Swift, S3, and HDFS.
- ▶ Secures data at rest by way of encryption.
- ▶ Provides secure deletion of data.
- ▶ Provides secure and audited access to personal data through authentication and authorization, and file audit logging.
- ▶ Meets the GDPR compliance requirements in accordance to EU GDPR Article 21 Section 1.

For more information about functions in IBM Spectrum Scale that support GDPR, see *IBM Spectrum Scale Functionality to Support GDPR Requirements*, [REDP-5489](#).

For more information about the KPMG assessment report for IBM Spectrum Scale version 5.0, see [this web page](#).

For more information about IBM Spectrum Scale security features, see *IBM Spectrum Scale Security*, [REDP-5426](#).

Introduction to IBM StoredIQ

This section describes IBM StoredIQ scalable analysis and the governance of unstructured data and its use cases.

What is IBM StoredIQ

IBM StoredIQ provides scalable analysis and governance of unstructured data across many different data sources, such as email servers, file shares, and collaboration sites. It enables companies to discover, analyze, and act on data for eDiscovery and storage optimization.

For more information about supported data sources, see [this website](#).

Providing an in-depth and in-place assessment of data across many data sources and hundreds of file types, scaling from terabytes to petabytes and giving unprecedented visibility into unstructured data, IBM StoredIQ helps organizations to make more informed business decisions. It addresses the challenges of data discovery and remediation of personal data. Together with IBM StoredIQ for Legal, it delivers an end-to-end platform that streamlines eDiscovery for legal stakeholders. IBM StoredIQ includes the following features:

- ▶ In-place data assessment that allows an organization to discover, analyze, and act on unstructured data without moving it to a repository or specialty application.
- ▶ Powerful data discovery capabilities that use regular expressions, text analytics, custom annotators, word lists, and metadata to first filter and then discover personal data across enterprise-size landscapes.
- ▶ 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.
- ▶ Fully audited action/remediation that supports many different policies, such as copy, delete, move, copy to a repository with retention or export

IBM StoredIQ use cases

IBM StoredIQ helps businesses understand hundreds of file types across many different unstructured data sources. It provides the following functions:

- ▶ **GDPR/Privacy**

Uncovering hidden personal information and taking remedial actions to delete it or move it to a more secure location.

- ▶ **Legal/eDiscovery**

Identifying and collecting data that is potentially relevant to a legal matter. Works closely with IBM StoredIQ for Legal to align legal and IT stakeholders.

- ▶ **Data Minimization**

ROT (Redundant, Obsolete, Trivial) data clean up. Identifying and deleting low business value or highly risky data and removing it from corporate networks.

Based on this unique set of functions, four primary use cases are available for IBM StoredIQ, which are shown in Figure 2 on page 5 and described in the following sections.

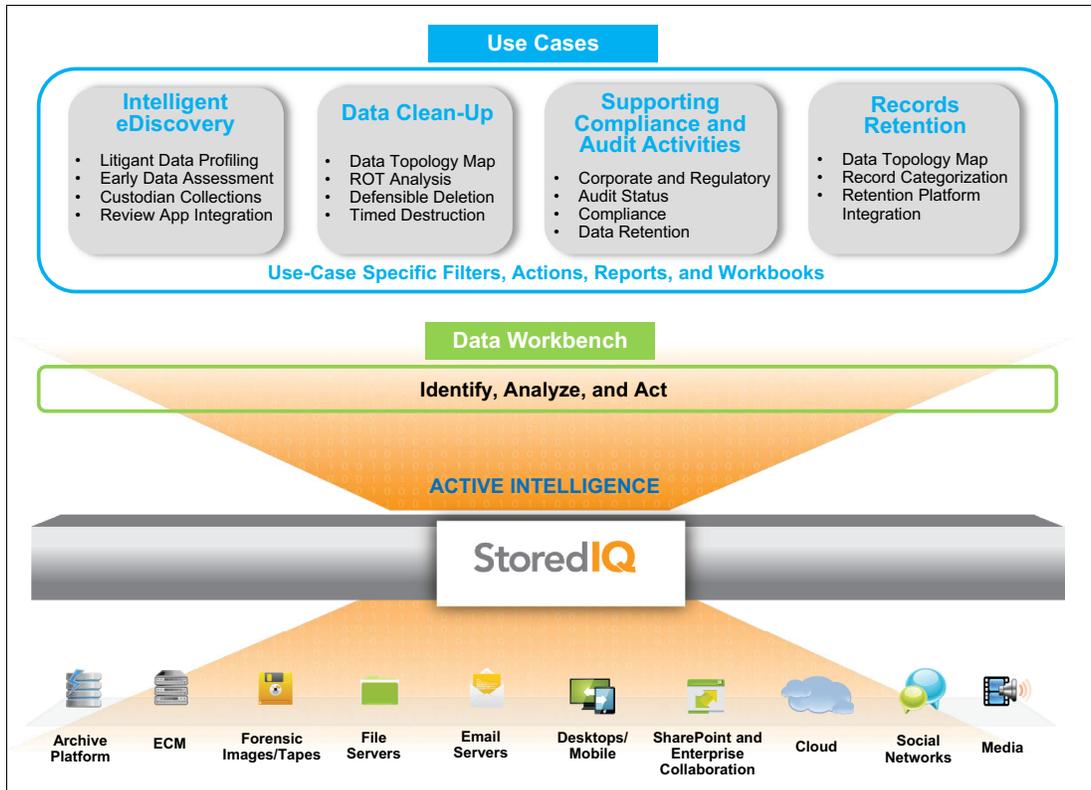


Figure 2 IBM StoredIQ use cases

Use Case: Intelligent eDiscovery

According to various studies, data capacity in enterprises is growing 40% - 60% year over year. This phenomenal data growth is the result of many factors, such as an explosion in unstructured data or content (email, documents, images, videos, and so on). Some of this unstructured data must be stored in accordance to regulatory requirements that continue to evolve and change.

Executive management wants to improve the efficiency of finding relevant data, including hidden or dark data. This process requires understanding and cataloging unstructured data, which can be spread all over the globe by scanning through the relevant data sources and storage systems.

StoredIQ Intelligent eDiscovery and advanced visualizations can show you what data is stored and find hidden and dark data. StoredIQ can connect to many different data sources and recognize more than 450 different file types. It also provides a connector toolkit (software development toolkits) that helps flexibly adopt new data sources and file types. This feature reduces the risk of not being able to respond to regulatory audits.

Use case: Data Clean-Up

Today, many organizations are challenged with task of identifying useful data among data that is redundant, less important, or obsolete. The ever-growing volume of data makes this challenge even bigger. After data is identified and understood in most deployments, only a subset of the overall data (approximately 20%) is found to be useful. The remaining data might be useful, but most of it likely is redundant, obsolete, or less important.

Executive management wants to improve the efficiency of the process that is used to identify and use useful data while cleaning up data storage from redundant, less important, and obsolete data. IBM StoredIQ provides various levels of filters that can be applied across the data sources. After these filters are applied, data that is classified as redundant, obsolete, or less important can be deleted or moved by IBM StoredIQ to a lower-cost storage if deletion is not authorized.

Use case: Supporting Compliance and Auditing activities

Sensitive data, such as personal information and confidential information and intellectual property, often is subject to regulations and underlies special protection requirements against data breaches and unauthorized access. Regulations can be internal or external to an organization; for example, external regulations, such as CCPA, HIPAA, BCBS239, Solvency II, and General Data Protection Rule (GDPR) for EU and its related countries.

Organizations are audited by authorities to check compliance regarding regulations. Auditors can use any data that is provided against the organization, including obsolete, redundant, and less important data. This issue poses a double risk to organizations: they must identify and secure sensitive data and remove access to redundant and less important data.

Executive management wants to reduce the risk of compliance exposure for sensitive data among all the known and unknown data sources within the organization. With IBM StoredIQ, organizations can make use of predefined filters (called *policies*) according to these regulations to identify sensitive data and establish a process for protecting it.

Use case: Records Identification and Retention

One of the strengths of IBM StoredIQ is that it supports an iterative approach to improve information governance. Applying retention policies to data and information that is subject for regulatory compliance is one of the key functions of information governance. IBM StoredIQ's automated classification enables organizations to quickly and accurately identify information that is subject for regulatory compliance, for remediation and ongoing compliance assurance.

Based on the corporate retention policies, executive management wants to specifically identify content that is spread across the organizations, which is required to be retained for a specific period. By using IBM StoredIQ to identify data that is outside the records repository, this process can be greatly improved. Data can be filtered, culled, and classified based on metadata and keywords by the integration with IBM Content Classification. Records are then properly identified, relocated, or dispositioned by using the rich set of features that is included in IBM StoredIQ.

IBM StoredIQ business benefits

IBM StoredIQ provides the following business benefits:

- ▶ **Reduced cost:** Discovering obsolete or redundant data results in reduced storage cost.
- ▶ **Reduced risk:** Discovering sensitive and personal data helps remediate that data.
- ▶ **Increased protection:** Valuable data assets are protected by using proper management and control.

For more information about IBM StoredIQ, see “Related publications” on page 22.

IBM StoredIQ with IBM Spectrum Scale: Harvesting and Protecting Data

Identifying and classifying data that is subject to regulatory compliance is one of the key challenges that many organizations are facing. For example, the EU GDPR requires organizations to identify and protect data with personal information. This challenge is multiplied if the data is across different islands of storage and perhaps in different regions and continents. This challenge can be addressed by consolidating data within a single highly scalable global namespace that is provided by IBM Spectrum Scale.

The identification of personal information that is stored in unstructured data requires intelligent analytic software. This capability is provided by IBM StoredIQ, which includes dedicated GDPR cartridges to identify personal information. IBM StoredIQ Cartridges are analytic plug-ins that contain more analysis logic that can be uploaded to IBM StoredIQ. Cartridges can contain analysis logic that is based on different technologies that range from simple regular expressions to full blown cognitive approaches, such as natural language processing (NLP). By adding a cartridge to the IBM StoredIQ, you enable it to find and index sensitive data in documents, thus making them searchable.

For example, a sensitive pattern cartridge for the General Data Protection Regulation (GDPR) can enable IBM StoredIQ to detect addresses, passport numbers, phone numbers, and other personal identifiers. These GDPR cartridges are delivered as samples and support the following types of personal data (for all European countries) by default (for more information, see the cartridge readme file). Depending on the cartridge, the following information can be analyzed within unstructured content:

- ▶ International Bank Account Numbers (IBAN)
- ▶ IDs:
 - Passport
 - National ID
 - Social Security Number (not for all countries)
 - Tax Number (not for all countries)
- ▶ Phone Numbers
- ▶ Email Addresses
- ▶ IP addresses
- ▶ Person Names (not for all countries)
- ▶ Locations (not for all countries)
- ▶ Organizations (not for all countries)
- ▶ Addresses (not for all countries)
- ▶ Dates (in different formats)

These cartridges can be extended or you can create your own cartridge. For more information, see [this website](#).

These GDPR cartridges can be downloaded from [IBM Fix Central](#).

In the next sections, we describe the integration of IBM StoredIQ and IBM Spectrum Scale, whereby:

- ▶ Unstructured data is in an IBM Spectrum Scale file system.

- ▶ IBM StoredIQ is configured to mount the file system that is provided by IBM Spectrum Scale by way of NFS and performs deep inspection and identification of personal data that is in the IBM Spectrum Scale file system. This process results in getting a CSV list of files that contain personal data (including the match that was found in that file with details) that must be appropriately handled according to the requirements (for example, GDPR).
- ▶ The identified files are secured by moving them into a portion of the IBM Spectrum Scale file system where the files are encrypted, which uses IBM Spectrum Scale security functions.

Note: Moving the identified files to a secure area in the IBM Spectrum Scale file system is used as an example. Depending on the legal requirement and the nature of the personal information, further actions might have to be performed for the identified files and its contents, such as anonymization of data and deletion of data or records.

Environment

As shown in Figure 3, an IBM Spectrum Scale cluster is configured with a file system that hosts mixed data with personal and non-personal information. This file system contains a directory that we call *email_data* where emails are stored. The emails that are stored in this directory are taken from a hypothetical email data research project. It contains data from approximately 150 users that is organized into folders. The directory *email_data* is exported by way of NFS and mounted by the server that is running IBM StoredIQ.

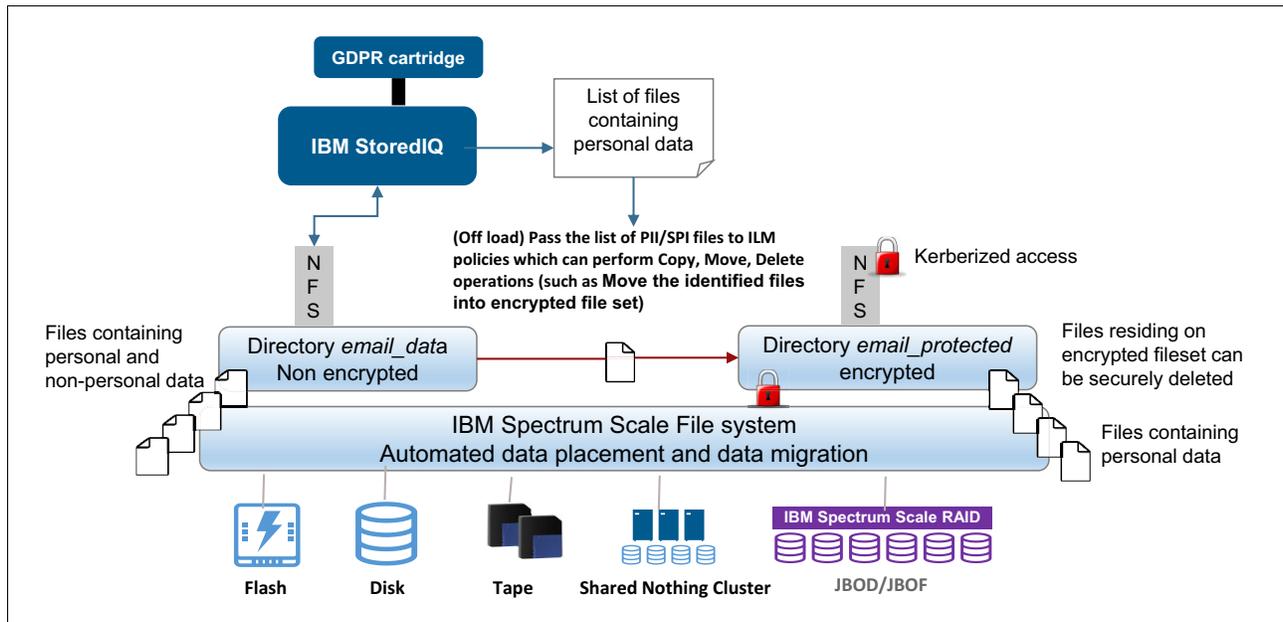


Figure 3 IBM StoredIQ inspecting and identifying unstructured data that is stored in IBM Spectrum Scale file system

IBM StoredIQ is deployed and the GDPR cartridges to identify personal information are loaded into IBM StoredIQ. In this example, we use IBM StoredIQ to identify files that contain phone numbers by using the `gdpr-basic` cartridge. It is possible to identify much more personal information, including person names, addresses, and IDs.

However, we want to keep this demonstration simple. The result of the inspection and analysis that is performed by IBM StoredIQ is a list of all file names that match the filters and rules of the GDPR cartridge.

The Spectrum Scale file system provides another directory that we call *email_protected*. All files that are stored in this directory are automatically encrypted by the IBM Spectrum Scale file encryption function. All files that are identified by the IBM StoredIQ GDPR cartridge are moved into this directory and can then be accessed only when the appropriate encryption key is provided. Also, files that are stored in this directory can be secured in motion by using kerberized NFS.

Configuring IBM Spectrum Scale

The IBM Spectrum Scale cluster that is used for this demonstration was configured with protocol nodes (NFS/SMB/Object). For more information about installing and configuring IBM Spectrum Scale with protocol nodes, see *Installing and deploying IBM Spectrum Scale with the installation toolkit*, which is available at this [IBM Knowledge Center web page](#).

One file system was created and mounted under the path name /gpfs0. In this file system, the directory /gpfs0/email_data represents our email_data directory. This directory is exported to the IBM StoredIQ server and mounted as volume later on. The directory /gpfs0/email_protected represents our email_protected directory and is configured to encrypt all files that are stored in this directory.

Complete the following steps:

1. Export the email_data directory as an NFS share. In Example 1, the email_data is an IBM Spectrum Scale file set that is in the path /gpfs0/email_data.

Example 1 Export the email_data directory as an NFS share

```
# mmnfs export add /gpfs0/email_data --client "*(Access_Type=RW,Protocols=3:4)"
mmnfs: The NFS export was created successfully

# mmnfs export list
Path                Delegations  Clients
-----
/gpfs0/email_data   NONE         *
```

Example 2 shows sample content that is stored in the file set.

Example 2 Sample email data set stored in the IBM Spectrum Scale file set

```
# ls /gpfs0/email_data/sq_fset/maildir
ermis-f mccarty-d mckay-b mclaughlin-e

# ls /gpfs0/email_data/sq_fset/maildir/ermis-f/
all_documents deleted_items discussion_threads inbox notes_inbox sent
sent_items _sent_mail

# ls /gpfs0/email_data/sq_fset/maildir/mccarty-d/
calendar contacts deleted_items inbox sent_items to_do

# ls /gpfs0/email_data/sq_fset/maildir/mclaughlin-e/
all_documents contacts discussion_threads gossett inbox notes_inbox
private_folders sent
_sent_mail calendar deleted_items eol__tagg greg_couch monthly_pma_s
prepays_structured_deals
security_request sent_items tasks
```

2. Configure the `email_protected` directory for encryption by activating an encryption policy. The `email_protected` directory is under the path `/gpfs0/email_protected` and is configured as an IBM Spectrum Scale file set. The integration with an external IBM SKLM server is not shown in this publication. We assume that the key manager was configured.

The encryption policy configuration for the directory `email_protected` is shown in the following example and it is stored in the `gpfs0.policy.txt` file. The encryption key that is used for encryption of all files that are stored in the `email_protected` directory includes the label `0bb9856f-50c3-3b9c-9c9d-a70a59d8de0f` and the key server is configured under the alias `sk1m1`:

```
RULE 'defaultplacement' SET POOL 'system'  
RULE 'encryptemail' set ENCRYPTION 'E1' FOR FILESET ('email_protected')  
RULE 'keyrule' ENCRYPTION 'E1' IS ALGO 'DEFAULTNISTSP800131A'  
KEYS ('0bb9856f-50c3-3b9c-9c9d-a70a59d8de0f:sk1m1')
```

The following command is used to activate this policy:

```
# mmchpolicy gpfs0 gpfs0.policy.txt
```

Configuring IBM StoredIQ

In this section, we show how we configured IBM StoredIQ with the cartridge `gdpr-basic` to identify files that are stored in the `email_data` directory that contain phone numbers. For this purpose, we upload the `gdpr-basic` cartridge into IBM StoredIQ and then mount the `email_data` directory by way of NFS as volume into the IBM StoredIQ server.

To configure IBM StoredIQ, start the IBM StoredIQ AppStack UI and complete the following steps:

1. Check that the data server is available. Click **Administrator console** → **Data Servers**.

The Data Server is a component of the IBM StoredIQ installation. It obtains the data from supported data sources and indexes it (see Figure 4).

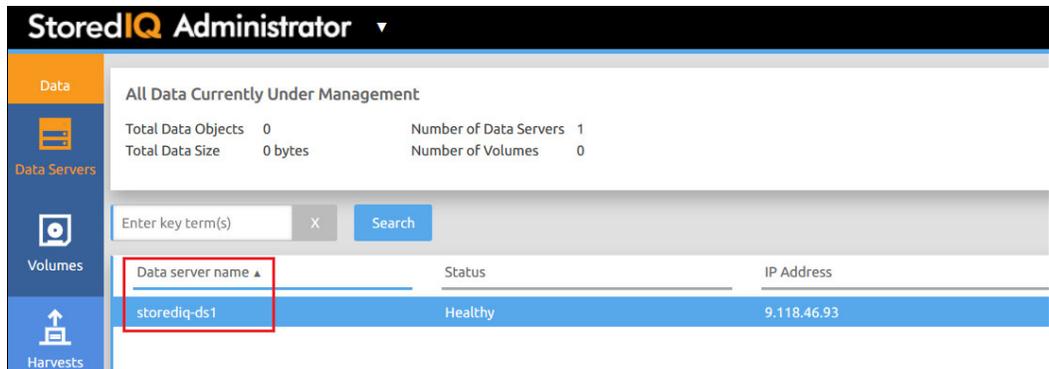


Figure 4 Displaying all data that is under management by IBM StoredIQ Data Servers

2. Upload GDPR cartridges `gdpr-basic`. On the Administrator console, click **Cartridges** → **Upload Cartridge**.

The GDPR Focused Data Discovery cartridges are designed to assist organizations to identify, comply with GDPR, and other personal data regulations. Different types of GDPR cartridges are available. In this demonstration (see Figure 5), we use the most basic cartridge type (named `gdpr-basic`).

Upload a cartridge to use in an analytics step-up.

Enter key term(s) X

Name	Cartridge name	Supported results	Status	Creation date
<code>gdpr-basic</code>	GDPRFocusedDataDiscoveryBasic	BankAccountNumber,PhoneNumber,EmailA...	Available	2018-09-12 4:30 PM

Figure 5 Uploading basic type cartridge for use in analytics step-up

Note: The cartridge `gdpr-basic` allows users to analyze the unstructured data based on regular expressions. This analyzation is sufficient to identify phone numbers. For more advanced analytics, such as the identification of person names, addresses, and other non-English language support, the cartridge `gdpr-advanced` can be used.

3. Add the NFS export of the directory `email_data` as a volume in IBM StoredIQ, as shown in Figure 6. On the Administrator console, click **Volumes**. Enable the indexing option **Include metadata for contained objects**, when required. You cannot enable the indexing option **Include content tagging and Full-Text index** because the full text index is generated during harvesting and step-up analytics.

Add Volume

Volume Type:

Source Type:

Server:

Assign To Data Server:

Volume Name:

Export:

Initial Directory:

Indexing Options

Include metadata for contained objects

Include content tagging and Full-Text Index

Figure 6 Adding the NFS export of the directory `email_data` as a volume in IBM StoredIQ

After the successful addition of a volume, it can be viewed in the Volume tab, as shown in Figure 7.

Note: Data objects are shown as “0” until the harvest is completed.

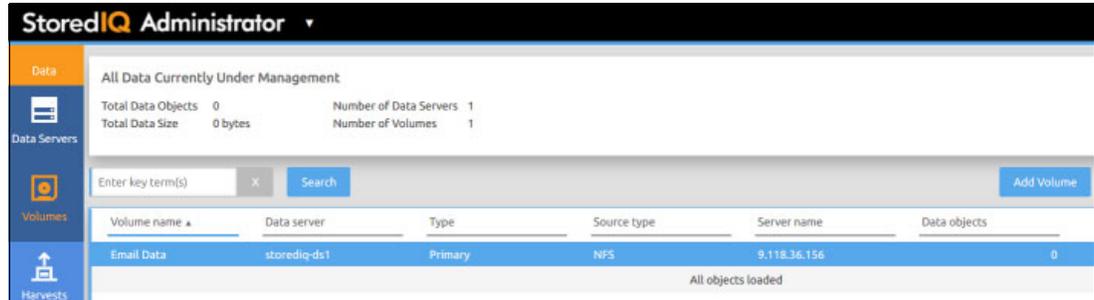


Figure 7 IBM StoredIQ Volume view

Harvesting volume with email data

In this section, we show how to configure IBM StoredIQ to harvest the email data that is stored in the IBM Spectrum Scale file system by using the GDPR cartridge. Harvesting essentially means that the data objects (files) of a volume are being indexed according to the index options selected for the data volume. The indexing options are shown in Figure 6 where the NFS export is added.

Note: It is recommended to prevent write access to the NFS export (directory *email_data*) during this operation. This process can be achieved by configuring the NFS export for read-only access.

Complete the following steps:

1. Start harvesting on the volume that contains the email data. On the Administrator console, click **Volumes** → **Harvest** (see Figure 8).



Figure 8 Starting the harvesting process on the volume containing the email data

2. Configure harvest settings to run a full harvest immediately, as shown in Figure 9.

Harvest Volume

Harvest Name
Harvest Email Data

Harvest Volume

Email Data

Schedule Harvest

Immediate - This harvest will be queued to run when you complete this dialog.

Schedule

On at UTC+05:30

Harvest Options

Incremental

Full

Figure 9 Selecting the settings to run a full harvest immediately

Harvest progress can be viewed on the Administrator console by clicking **Harvests** → **Current**, as shown in Figure 10.

StoredIQ Administrator

Select a harvest instance below.

Enter key term(s) Search

Name	Type	Complete	Start time	Est. completion
Harvest Email Data	Full harvest	0.00%	2018-10-02 12:13 PM	Calculating..

Figure 10 Selecting Harvests and Current to view progress

After the process is complete, the harvest job (see Figure 11) can be viewed on the Administrator console by clicking **Harvests** → **Completed**.

Select a harvest instance below.

Enter key term(s) Search

Name	Type	Start time	End time	Total time	Owner
Harvest Email Data	Full harvest	2018-10-02 12:13 PM	2018-10-02 12:18 PM	4m:41s	super admin

Figure 11 Completed harvest job

Analyzing email data

In this section, we demonstrate how to create a user-defined infoset that is based on the harvested data that is being analyzed by step-up analytics by using the `gdpr-basic` cartridge. An infoset consists of a set of data objects from one or more volumes that match certain filters.

Complete the following steps:

1. Create a Step-up Analytics action that is named `GDPR_BASIC` by using the `gdpr-basic` cartridge, as shown in Figure 12.



Figure 12 Creating Step-up Analytics action

2. Select the `gdpr-basic` cartridge that was loaded before (see Figure 13).

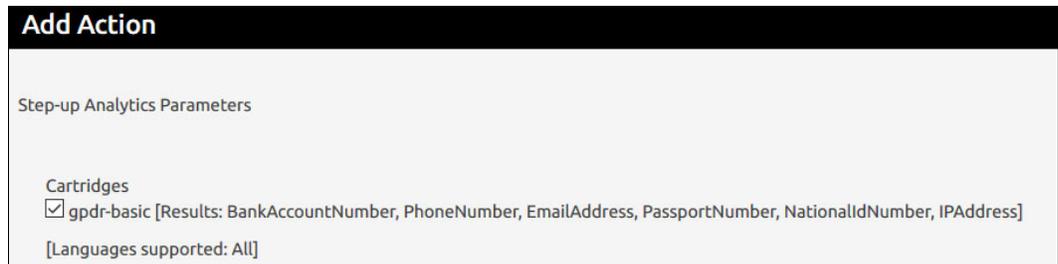
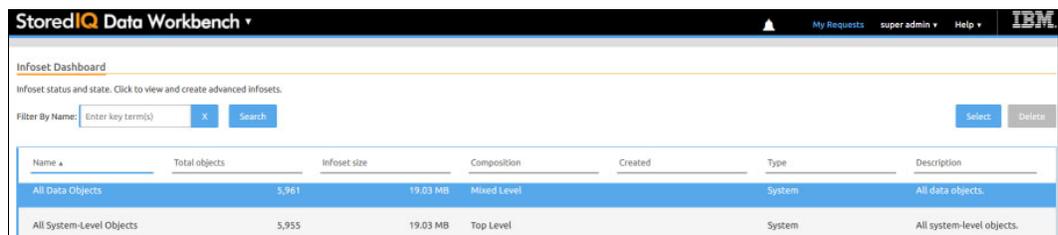


Figure 13 Selecting the `gdpr-basic` cartridge

3. Switch to the Data Workbench console, which lists the default system infosets. Select **All Data Objects** from the system infosets. Create a user-defined infoset by clicking **Select** (see Figure 14).



Name	Total objects	Infoset size	Composition	Created	Type	Description
All Data Objects	5,061	19.03 MB	Mixed Level		System	All data objects.
All System-Level Objects	5,955	19.03 MB	Top Level		System	All system-level objects.

Figure 14 Selecting the All Data Objects from the system infosets

4. In this example, the new user-defined infoset is named `ALL_USER_DATA` and it is based on the system infoset `All Data Objects` because this infoset contains the indexed (harvested) data for the volume `email_data`.

Note: A user define infoset can also be created for a particular set of volumes. Because we have only one indexed volume, we can use the indexed data from the system infoset All Data Objects.

Figure 15 shows the Filter Code that uses the default filter codes is sufficient for this exercise.

```
att:/Library/Attributes/System metadata/Object category' = any('Collaboration', 'Computer programming', 'Container files and archives', 'Email', 'Images', 'Miscellaneous', 'Multimedia', 'Presentation', 'Spreadsheets', 'System files', 'Word processing') and att:/Library/Attributes/System metadata/Volume type' is any('nfsv2', 'nfsv3')
```

Figure 15 Filter Code

Figure 16 shows creating a user-defined infoset that is used to run the analytics.

New Infoset Name:
ALL_USER_DATA

Description:
Enter infoset description

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 16 Creating a user-defined infoset ALL_USER_Data

Figure 17 shows the status of the user-defined infoset being created.

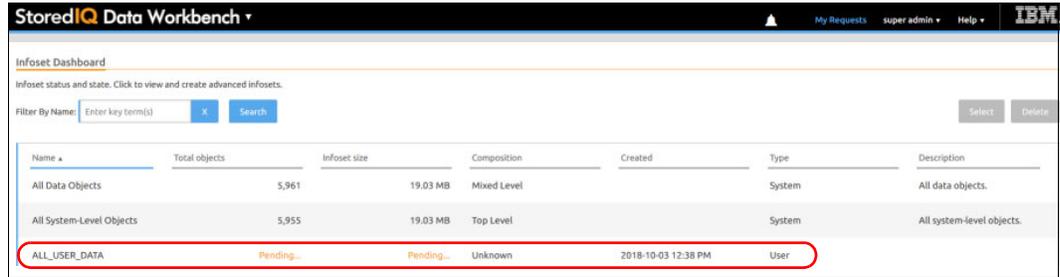


Figure 17 Infoset dashboard showing the status of the new infoset being created

5. Select the newly created infoset ALL_USER_DATA (see Figure 18) for running Step-up analytics.

Note: The new user-defined infoset ALL_USER_DATA includes the same number of objects as in the system infoset All Data Objects because it is a copy

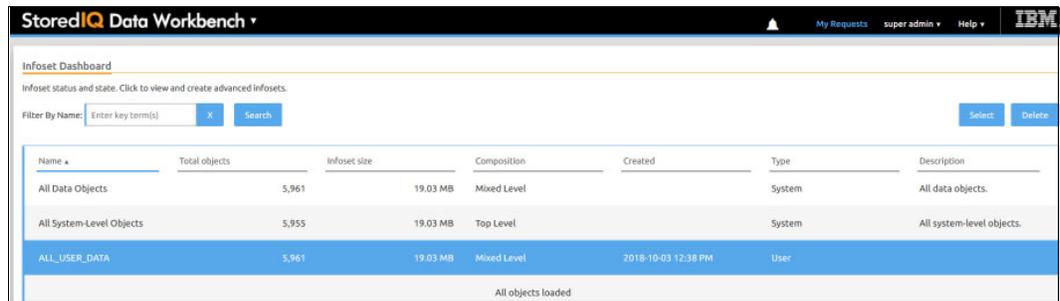


Figure 18 Infoset Dashboard listing all the infosets with ALL_USER_DATA as a selection.

Figure 19 shows the details of the selected ALL_USER_DATA infoset, which indicates different information, such as the number of files and the total size of the infoset.

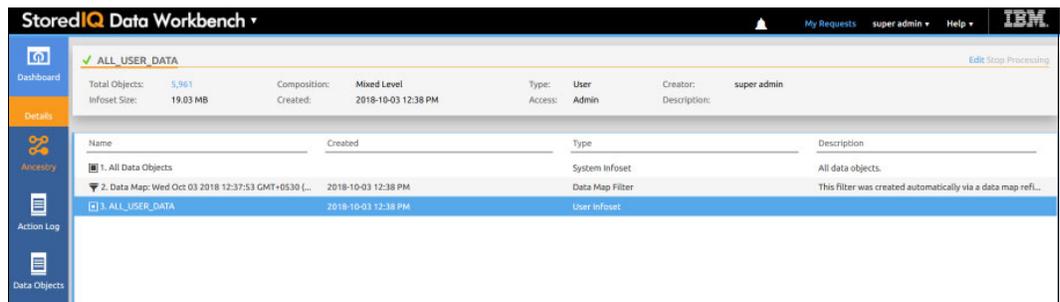


Figure 19 Detail of the selected ALL_USER_DATA infoset

6. Run the created Step-up-Analytics (GDPR_BASIC) action on the selected user infoset (see Figure 20 on page 17).

Note: Step-up Analytics runs a selected cartridge on an infoset (in this case, the cartridge gdpr-basic is used). IBM StoredIQ examines all documents in the infoset, applies the analytics that is contained in the cartridge to the data objects (files), and then stores the analysis results in the IBM StoredIQ index.

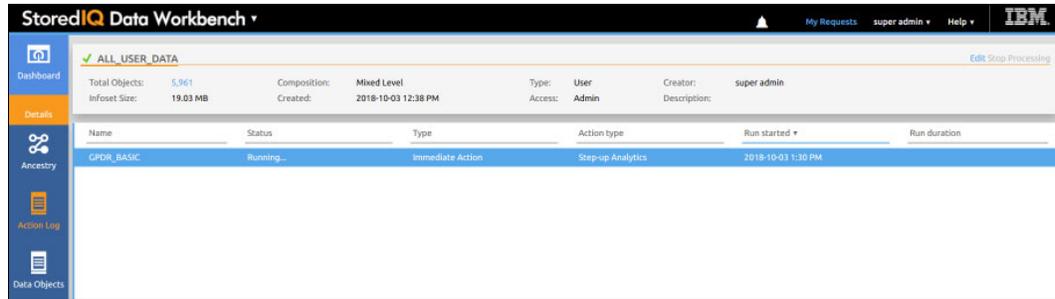


Figure 20 GDPR_BASIC step-up analytics being run on ALL_USER_DATA infoset

Figure 21 shows the final status of the step-up analytics.

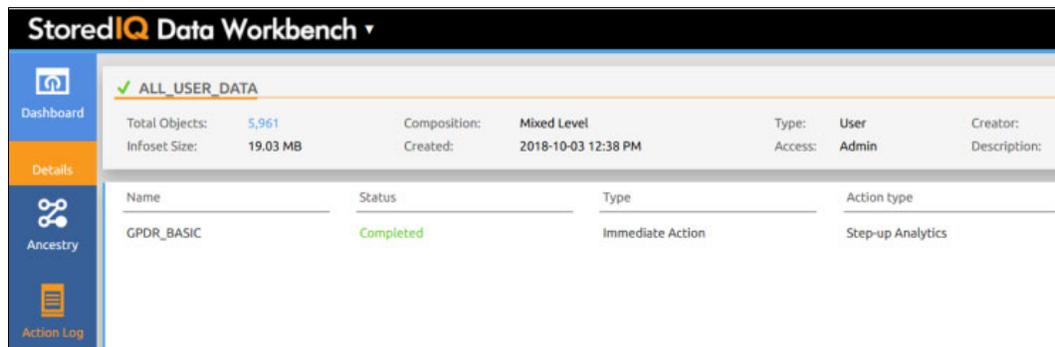


Figure 21 GDPR_BASIC step-up analytics successfully completed on ALL_USER_DATA infoset

Creating a list of files containing phone numbers

In this section, we describe how to create a user-defined infoset that is named DATA_WITH_PHONENUMBER by searching for phone numbers in previously analyzed infoset ALL_USER_DATA. The new infoset contains all data objects that include phone numbers.

Complete the following steps:

1. To find a list of files that contain phone numbers within the user-defined infoset ALL_USER_DATA that was created before using the cartridge gdpr-basic, create a filter by using PhoneNumber as search term. Save this filter as GDPR_PhoneNumber (see Figure 22).

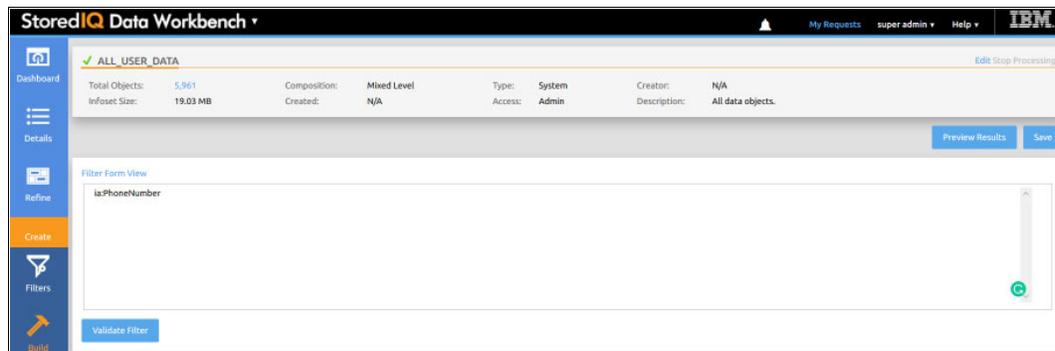
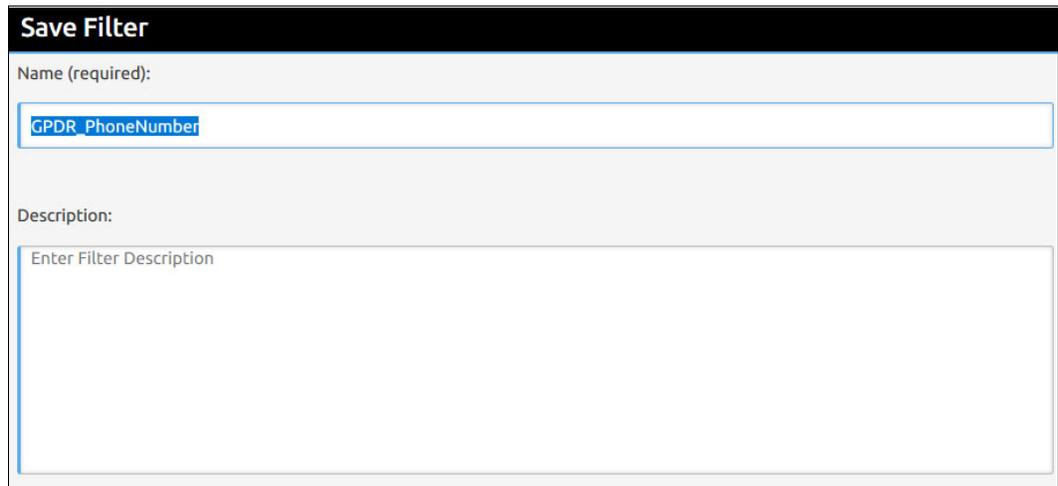


Figure 22 Selecting PhoneNumber as the filter to be applied on the ALL_USER_DATA infoset

Figure 23 shows that the filter must be saved after the filter is selected.



Save Filter

Name (required):

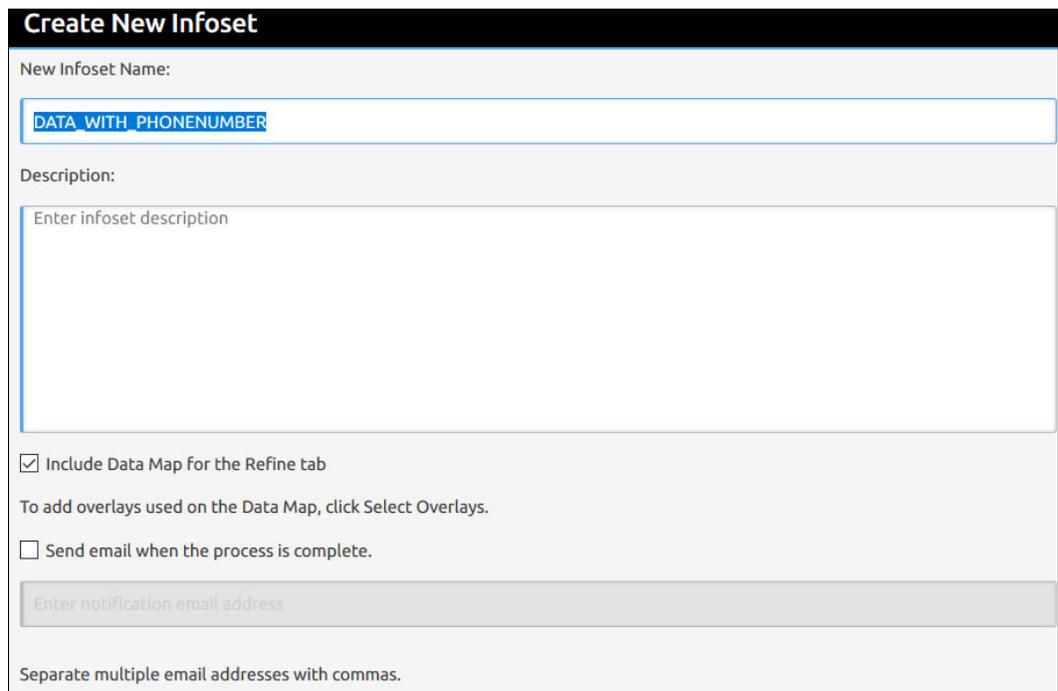
GDPR_PhoneNumber

Description:

Enter Filter Description

Figure 23 Saving the filter was selected

2. Create a user-defined infoSet (see Figure 24) with the previously created filter GDPR_PhoneNumber, which is used to store the filter results. The new infoSet is named DATA_WITH_PHONENUMBER.



Create New InfoSet

New InfoSet Name:

DATA_WITH_PHONENUMBER

Description:

Enter infoSet description

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 24 Creating a user-defined infoSet called DATA_WITH_PHONENUMBER

Figure 25 shows how to add the overlays (ALL_USER_DATA infoset and GPDR_PhoneNumber filter) that is used for creating the infoset.

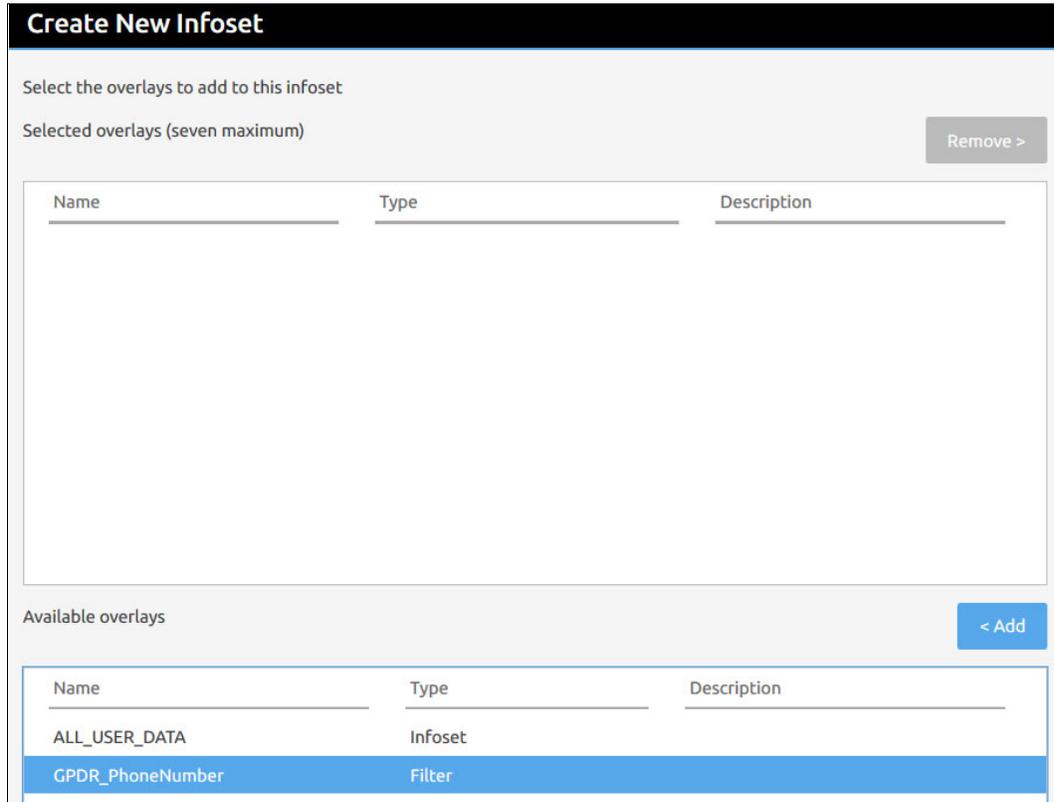


Figure 25 Adding the overlays for creating the new user-defined Infoset

- Wait for the infoset DATA_WITH_PHONENUMBER to be created. The pending status is shown in Figure 26.

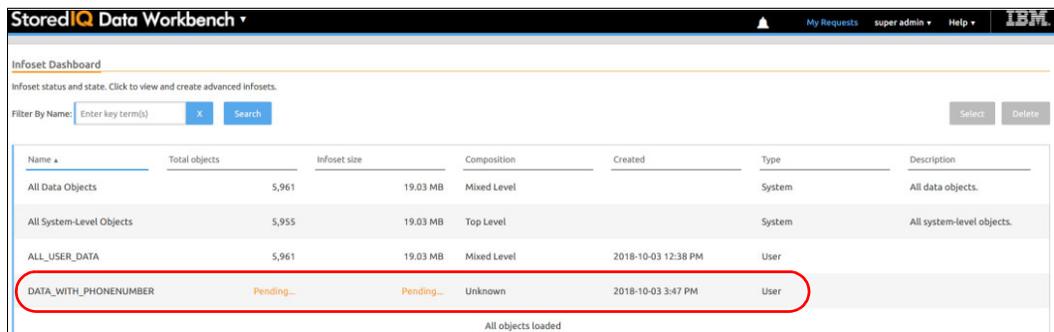


Figure 26 Infoset dashboard showing the status of the new Infoset being created

- When the infoset creation process is complete, the status is displayed in the Infoset Dashboard, as shown in Figure 27 on page 20.

Note: The number of data objects in the new infoset DATA_WITH_PHONENUMBER is less than the number of data objects that are analyzed in infoset ALL_USER_DATA. The difference is because the new infoset contains only data objects (files) where phone numbers are found.

Name	Total objects	Infoset size	Composition	Created	Type	Description
All Data Objects	5,961	19.03 MB	Mixed Level		System	All data objects.
All System-Level Objects	5,955	19.03 MB	Top Level		System	All system-level objects.
ALL_USER_DATA	5,961	19.03 MB	Mixed Level	2018-10-03 12:38 PM	User	
DATA_WITH_PHONENUMBER	79	400.35 KB	Top Level	2018-10-03 3:47 PM	User	

Figure 27 Infoset dashboard showing the newly created Infoset DATA_WITH_PHONENUMBER

- To view list of data objects (files) that contains GDPR information (phone number), in the Workbench console, click **Details** → **Data Objects** (see Figure 28).

File name	File size	File path	Created	Last modified
163.	13.67 KB	sq_fset/maildir/fermis-f/all_documents	2018-10-02 11:43 AM	2018-10-02 11:43 AM
158.	13.19 KB	sq_fset/maildir/fermis-f/all_documents	2018-10-02 11:43 AM	2018-10-02 11:43 AM
130.	1.27 KB	sq_fset/maildir/mclaughlin-e/all_documents	2018-10-02 11:42 AM	2018-10-02 11:42 AM
105.	2.36 KB	sq_fset/maildir/mclaughlin-e/discussion_thr...	2018-10-02 11:42 AM	2018-10-02 11:42 AM
495.	1.04 KB	sq_fset/maildir/mclaughlin-e/discussion_thr...	2018-10-02 11:42 AM	2018-10-02 11:42 AM
189.	2.78 KB	sq_fset/maildir/fermis-f/notes_inbox	2018-10-02 11:43 AM	2018-10-02 11:43 AM
111.	1.63 KB	sq_fset/maildir/fermis-f/inbox	2018-10-02 11:43 AM	2018-10-02 11:43 AM
244.	3.08 KB	sq_fset/maildir/fermis-f/all_documents	2018-10-02 11:43 AM	2018-10-02 11:43 AM
110.	3.12 KB	sq_fset/maildir/mclaughlin-e/all_documents	2018-10-02 11:42 AM	2018-10-02 11:42 AM
240.	2.79 KB	sq_fset/maildir/fermis-f/all_documents	2018-10-02 11:43 AM	2018-10-02 11:43 AM
683.	1.63 KB	sq_fset/maildir/fermis-f/inbox	2018-10-02 11:43 AM	2018-10-02 11:43 AM

Figure 28 Panel showing list of all the data objects (files) in DATA_WITH PHONENUMBER Infoset

Double-click the file that is presented under data objects to open the object in Data viewer. The content is displayed (see Figure 29) that was found in the file along with the related personal information.

Name: 495. Path: sq_fset/maildir/mclauidry-e/discussion_threads Size: 1.04 KB

< Previous page | 1 | Next page > (1 of 1 available)

I'm glad we ran into each other at the city's financial seminar. Good to catch up on old times as well as discuss possible new opportunities.

I am out of town next week so how about planning on getting together the following week the afternoon of Wednesday the 15th?

Chris V. McLauwdry

VP - Financial Trading Risk Mgmt
Corporation Z North America

Voice: xxx-xxx-xxxx
email: CVM@xxxxZ.com

Close

Figure 29 Data Object Viewer showing the content of the selected file

You can generate multiple reports that are related to the infoset by clicking **Reports** in the Workbench console (see Figure 30). Generated reports can be exported for viewing outside of the Workbench console.

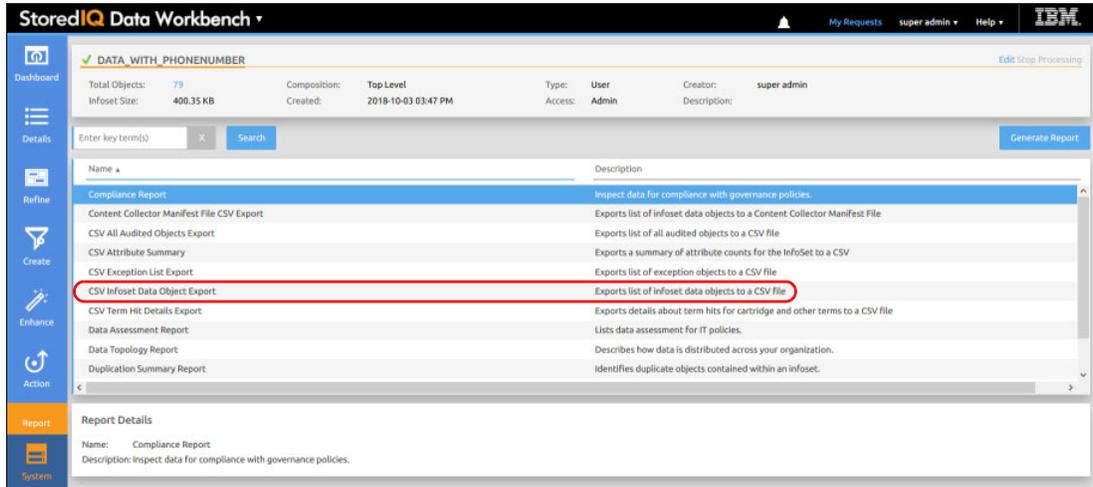


Figure 30 Panel showing reports that can be generated on the DATA_WITH_PHONENUMBER infoset

Encrypting identified files

The files in the email_data directory that contain personal information that are identified by analyzing the GDPR cartridge must be moved to the email_protected directory. The exported file list can be parsed in a custom shell script; for example, where the path and file names are extracted and used as source for the copy process. The files are encrypted by the IBM Spectrum Scale encryption feature when they are placed in the email_protected directory.

The files are copied from the email_data directory to the email_protected directory and must be removed from the email_data directory to avoid any possible access to the unencrypted version of the file.

In addition to the encryption feature, customers can consider the use of the file audit logging feature of IBM Spectrum Scale to track changes to the files in the email_protected directory, and the immutable file features to prevent deleting the files.

Note: File audit logging is enabled at the file system level. If file audit logging is not wanted for the entire file system, a separate file system can be used for the email_protected directory.

For more information about how to enable and configure the IBM Spectrum Scale encryption, file audit logging, immutability, protocols (NFS), and securing data in flight with Kerberos, see [IBM Knowledge Center](#).

Note: For more information about the KPMG assessment report and certificate of the immutability function of IBM Spectrum Scale Version 5.0 in accordance to US SEC17a-4f, EU GDPR Article 21 Section 1, German and Swiss laws and regulations, see the following resources:

- ▶ [Certificate](#)
- ▶ [Full assessment report](#)

Conclusion

This IBM Redpaper publication demonstrated the identification and protection of data subject to compliance with EU GDPR by using the combination of IBM Spectrum Scale storage and IBM StoredIQ analytics. The files that are subject to regulatory compliance were stored as unstructured data in an ordinary IBM Spectrum Scale file system. IBM StoredIQ was configured with a GDPR cartridge that contains filters and rules that are used to identify files, including phone numbers.

The identified files were moved to a secure area within the IBM Spectrum Scale file system that allows for data encryption. Encrypted data can be accessed only by users that are authorized to obtain the encryption key (other users cannot read the content of these files). Therefore, this solution helps to protect personal information from unauthorized access. At the same time, it allows other business processes to work with these files and even delete them when required.

The example that is presented in this publication demonstrated was simplified to provide a better understanding about how IBM Stored IQ and IBM Spectrum Scale can be used for data governance. The analytics of unstructured data can be much more advanced where not only phone numbers but names, addresses, IDs, and other personal information is identified; for example, by using the `gdpr-advanced` cartridge. Other tools, such as IBM StoredIQ Insight®, can be used to provide a professional view of the identified data to derive better business decisions.

Note: Clients are responsible for ensuring their own compliance with various laws and regulations, including the European Union General Data Protection Regulation. Clients are solely responsible for obtaining advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulations that may affect the clients' business and any actions the clients may need to take to comply with such laws and regulations.

The products, services, and other capabilities described herein are not suitable for all client situations and may have restricted availability. IBM does not provide legal, accounting or auditing advice or represent or warrant that its services or products will ensure that clients are in compliance with any law or regulation.

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics covered in this paper.

The following IBM Redbooks® publications provide additional information about the topic in this document. Some publications that are referenced in this list might be available in softcopy only:

- ▶ *IBM Spectrum Scale Immutability Introduction, Configuration Guidance, and Use Cases*, REDP-5507:
<http://www.redbooks.ibm.com/abstracts/redp5507.html>
- ▶ *IBM Spectrum Scale Security*, REDP-5426:
<http://www.redbooks.ibm.com/abstracts/redp5426.html>
- ▶ *IBM Spectrum Scale Functionality to Support GDPR Requirements*, REDP-5489:
<http://www.redbooks.ibm.com/abstracts/redp5489.html>

- ▶ IBM StoredIQ Introduction and Planning Considerations, REDP-5315:

<http://www.redbooks.ibm.com/abstracts/redp5315.html>

You can search for, view, download, or order these documents and other Redbooks, Redpapers™, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

The following web pages are helpful for more information:

- ▶ Accelerate your GDPR Unstructured Data Discovery with IBM StoredIQ Cartridges:

<https://www.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=06012606USEN>

- ▶ Deploying and configuring IBM StoredIQ:

<https://ibm.biz/Bd2sVf>

- ▶ IBM StoredIQ supported data sources:

<https://ibm.biz/Bd2sVP>

- ▶ IBM StoredIQ supported file types:

<https://ibm.biz/Bd2sVM>

- ▶ IBM StoredIQ creating custom cartridges:

<https://ibm.biz/Bd2sVv>

- ▶ IBM StoredIQ cartridges can be downloaded from IBM Fixcentral:

<https://www.ibm.com/support/fixcentral>

- ▶ IBM Spectrum Scale Knowledge Center:

<https://ibm.biz/Bdinhb>

- ▶ Installing and deploying IBM Spectrum Scale with the installation toolkit:

<https://ibm.biz/Bd2sVm>

IBM StoredIQ YouTube videos

For more information about IBM StoredIQ, see the following YouTube videos:

- ▶ Find Confidential Information with IBM StoredIQ:

https://www.youtube.com/watch?time_continue=4&v=TpXjytT4GV0

- ▶ Finding Important Business Documents with IBM StoredIQ:

https://www.youtube.com/watch?time_continue=14&v=s80b7sfSJ4Q

- ▶ Finding Personal Data with IBM StoredIQ:

https://www.youtube.com/watch?v=Ko_kRXz84_I

- ▶ Reduce Dark Data Risk With IBM StoredIQ:

https://www.youtube.com/watch?v=QE4PG_ZSX2w&t=376s

Testing IBM StoredIQ

Want to test drive IBM StoredIQ for yourself? Try an interactive demo of IBM StoredIQ that includes a quick tutorial to introduce you to the product. In this demonstration, you use IBM StoredIQ to find documents that contain the word “Confidential” and that were not modified in 5 years. This task might be part of a larger data cleanup effort or a step in a data migration process to meet new compliance rules. For more information, see [this website](#).

Delving deeper into IBM StoredIQ

For more information about the capabilities of IBM StoredIQ, see the [Information Lifecycle Governance - IBM StoredIQ - Data Intelligence](#) tutorials.

Authors

This paper was produced by a team of specialists from around the world working at IBM Redbooks, Tucson Center.

Sandeep R Patil is a Senior Technical Staff Member who works as a Storage Architect with IBM System Labs. He has over 18 years of product architecture and design experience. Sandeep is an IBM Master Inventor, an IBM developerWorks® Master Author, and a member of the IBM Academy of Technology. Sandeep holds a Bachelor of Engineering (Computer Science) degree from the University of Pune, India. He is recognized and listed by Wikipedia in the World Wide Prolific Inventors list.

Sasikanth Eda is a Software Engineer with the IBM Spectrum Scale development team. He works for the integration of OpenStack with IBM Spectrum Scale, focusing on Swift object, Cinder block, and Manila file storage components of OpenStack. Sasi holds a Masters degree in Microelectronics from I2IT, Pune, India.

Atul V Gore is an Information Technology Professional with IBM Enterprise Content Management Lab Services and has a total of 22 years of IT experience. He is IBM FileNet® P8 certified and is an expert in IBM Content Manager Enterprise Edition who successfully designed, installed, configured, maintained, and supported various IBM Enterprise Content Management customer solutions worldwide. He holds a Masters degree in Computer Science from the University of Pune, India.

Nils Haustein is a Senior Technical Staff Member at IBM Systems group and is responsible for designing and implementing backup, archiving, file, and Object Storage solutions in EMEA. He co-authored the book *Storage Networks Explained*. As a leading IBM Master Inventor, he has created more than 160 patents for IBM and is a respected mentor for the technical community worldwide.

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

Larry Coyne
IBM Redbooks, Tucson Center

Jayanth Gangadhar
Ashwin Kumar
Vivek Venkatanarasaiah
ECM Lab Services, IBM Hybrid Cloud, IBM India

Fred Stock
Carl Zeite
IBM Systems

Dirk Jahn
IBM StoredIQ, Technical Leader for Europe

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 IBM Redbooks 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.

Find out 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 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 US. This material might 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, MD-NC119, Armonk, NY 10504-1785, US

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 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 provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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 actual people or business enterprises 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.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, 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” at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®	IBM Elastic Storage™	Redbooks®
developerWorks®	IBM Spectrum™	Redpaper™
FileNet®	IBM Spectrum Scale™	Redpapers™
IBM®	IBM Z®	Redbooks (logo)  ®
IBM Cloud™	Insight®	StoredIQ®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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-5525-00

ISBN 0738457396

Printed in U.S.A.

Get connected

