IBM Spectrum Discover
Version 2.0.1

Concepts, Planning, and Deployment Guide

SC27-9601-05

Note

Before using this information and the product it supports, read the information in "Notices" on page 145.

Edition notice

This edition applies to version 2 release 0 modification 1 of the following product, and to all subsequent releases and modifications until otherwise indicated in new editions:

- IBM Spectrum Discover ordered through Passport Advantage (product number 5737-I32)
- IBM Spectrum Discover ordered through AAS/eConfig (product number 5641-SG1)

IBM® welcomes your comments; see the topic "How to send your comments" on page ix. When you send information to IBM, you grant IBM a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Figures

Tables

About this information

IBM Spectrum® Discover is metadata-driven management system for large scale file and object environments. IBM Spectrum Discover maintains a real-time metadata repository for large scale enterprise storage environments. Metadata can be searched, enhanced, discovered, and leveraged for data processing using built-in or custom agents.

Which IBM Spectrum Discover information unit provides the information you need?

The IBM Spectrum Discover library consists of the information units listed in Table 1 on page ix.

Table 1. IBM Spectrum Discover library information units

Information unit	Type of information	Intended users
IBM Spectrum Discover: Concepts, Planning, and Deployment Guide	This information unit provides information about the following topics: • Product Overview • Planning • Deploying and configuring	Users, system administrators, analysts, installers, planners, and programmers of IBM Spectrum Discover.
IBM Spectrum Discover: Administration Guide	This information unit provides information about administration, monitoring, and troubleshooting tasks.	Users, system administrators, analysts, installers, planners, and programmers of IBM Spectrum Discover.
IBM Spectrum Discover: REST API Guide	This information unit provides information about the following topics: • IBM Spectrum Discover REST APIs • Endpoints for working with a DB2 warehouse • Endpoints for working with policy management • Endpoints for working with connection management • Action agent management using APIs • RBAC management using APIs	Users, system administrators, analysts, installers, planners, and programmers of IBM Spectrum Discover.

Prerequisite and related information

For updates to this information, see IBM Spectrum Discover in IBM Knowledge Center (https://www.ibm.com/support/knowledgecenter/SSY8AC).

How to send your comments

You can add your comments in IBM Knowledge Center. To add comments directly in IBM Knowledge Center, you need to log in with your IBM ID.

You can also send your comments to ibmkc@us.ibm.com.

Summary of changes

This topic summarizes changes to the IBM Spectrum Discover licensed program and the IBM Spectrum Discover library. Within each topic, these markers ([ ]) surrounding text or illustrations indicate technical changes or additions that are made to the previous edition of the information.

Summary of changes
for IBM Spectrum
Discover version 2.0.1
as updated, August 2019

This release of the IBM Spectrum Discover licensed program and the IBM Spectrum Discover library includes the following improvements. All improvements are available after an upgrade, unless otherwise specified.

Deployment

You can create and scan IBM Spectrum Discover S3-compliant object storage data source connections. There are also updates on to how you can create and scan IBM Spectrum Discover and IBM Cloud™ Object Storage data connections.

Open virtual appliance (OVA)

The IBM Spectrum Discover software is available as an OVA (open virtualization appliance) file. You can deploy it on your VMware ESXi server by using the VMware vSphere Client.

Support for NFS sources

An NFS (Network File System) volume allows an existing NFS share to be mounted into your pod.

Creating content search policies

You can enrich metadata through content inspection of source data using the built-in **CONTENTSEARCH** agent. To use this function, you define regular expressions to search for and create policies that use the search results.

Hints and tips for using content-search policies

Running a content search policy on a set of documents has several steps, including retrieving the document, formatting it as text, if necessary, and searching the document. These are suggestions for best practices.

GUI changes

Ten regular expressions are included with this version, making it easier to create policies for such things as credit card formats, emails, and IPV4-addresses. Also, numerous Changes have been made to the GUI.

REST API changes

Documentation about the following REST API endpoints has been added:

```
/policyengine/v1/tags: GET
/policyengine/v1/tags/: POST
/policyengine/v1/tags/: PUT
/policyengine/v1/tags/: DELETE
```

REST APIs Documentation has been updated to provide information on new REST API features and functions. For more information, see the topic */policyengine/v1/tags: GET* in the *IBM Spectrum Discover: REST API Guide*.

[The **Collection Admin** role is available for administrators to:

- Update, and delete policies for the collections that they administer
- View, update, and delete data user policies for the collections that they administer

For more information, see the topic *Managing user access*.

Important: The **Collection Admin** role is available as a technical preview in the 2.0.1.1 release. For limitations on the usage of the **Collection Admin** role, see the *IBM Spectrum Discover Release Notes*.

Chapter 1. Product overview

Introduction to IBM Spectrum Discover

Companies need the ability to use unstructured data to meet their business priorities.

IBM Spectrum Discover is a modern metadata management software that provides data insight for petabyte-scale unstructured storage. The software easily connects to IBM Cloud Object Storage and IBM Spectrum Scale to rapidly ingest, consolidate, and index metadata for billions of files and objects.

IBM Spectrum Discover provides a rich metadata layer that enables storage administrators, data stewards, and data scientists to efficiently manage, classify, and gain insights from massive amounts of unstructured data. It improves storage economics, helps mitigate risk, and accelerates large-scale analytics to create competitive advantage and speed critical research.

Many companies face significant challenges to manage unstructured data. Unstructured data or unstructured information is defined as information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Some difficult challenges that companies face include:

- Pinpointing and activating relevant data for large-scale analytics.
- Lacking the fine-grained visibility needed to map data to business priorities.
- Removing redundant, trivial, and obsolete data.
- Identifying and classifying sensitive data.

Benefits of IBM Spectrum Discover

IBM Spectrum Discover can help you manage your unstructured data by reducing the data storage costs, uncovering hidden data value, and reducing the risk of massive data stores. See Table 2 on page 1.

Table 2. Benefits of IBM Spectrum Discover

Optimize - Improve storage usage	**Analyze - Uncover hidden data value**	**Govern - Mitigate risk and improve data quality**
Decreases storage capital expenditure (CaPex) by facilitating data movement to colder, cheaper storage.	Accelerates data identification for large-scale analytics.	Helps ensure that data is compliant with governance policies.
Increases storage efficiency by eliminating redundant data.	Operationalize tasks to reduce the burden of data preparation.	Helps reduce risk that is hidden in unstructured data stores.
Reduces storage operating expenditure (OpEx) by improving storage administrator productivity.	Orchestrates the ML/DL and Platform Symphony® MapReduce process.	Speeds the investigation into potentially fraudulent activities.

IBM Spectrum Discover architecture

IBM Spectrum Discover is an extensible platform that provides Exabyte scale data ingest, data visualization, data activation, and business-oriented data mapping.

Exabyte-scale data ingest

- Scan billions of files and objects in a day

- Real-time event notifications
- Automatic indexing

Data Visualization

- Fast queries of billions of records
- Multi-faceted search
- Drilldown Dashboard

Data Activation

- Action Agent SDK
- Extensible Architecture
- Solution blueprints

Business-oriented data mapping

- System-level data tagging
- Contextual data tagging
- Policy-driven work flows

Figure 1 on page 2 shows an example of the IBM Spectrum Discover architecture.

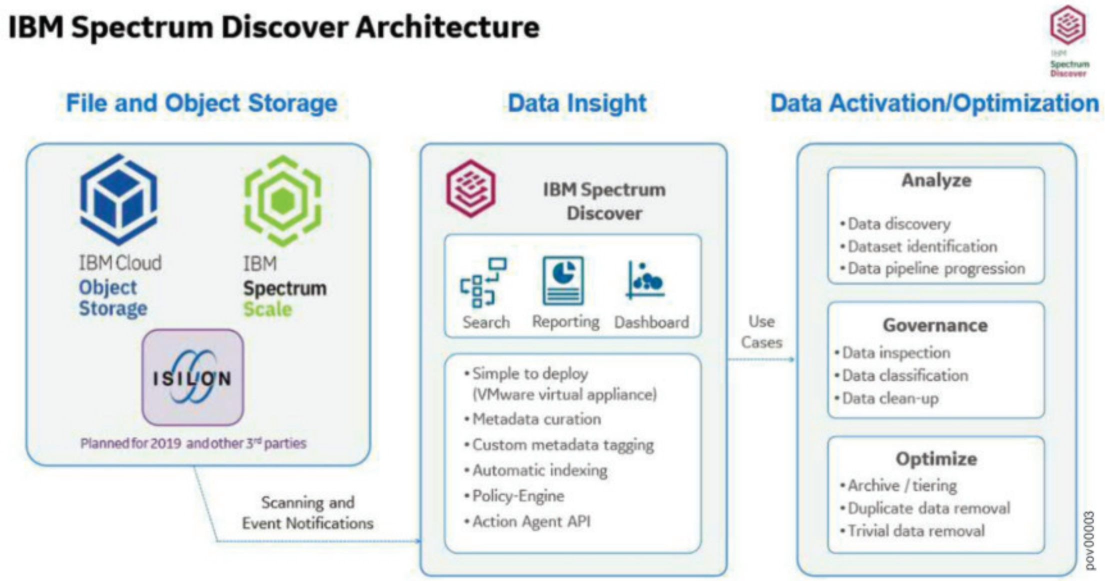
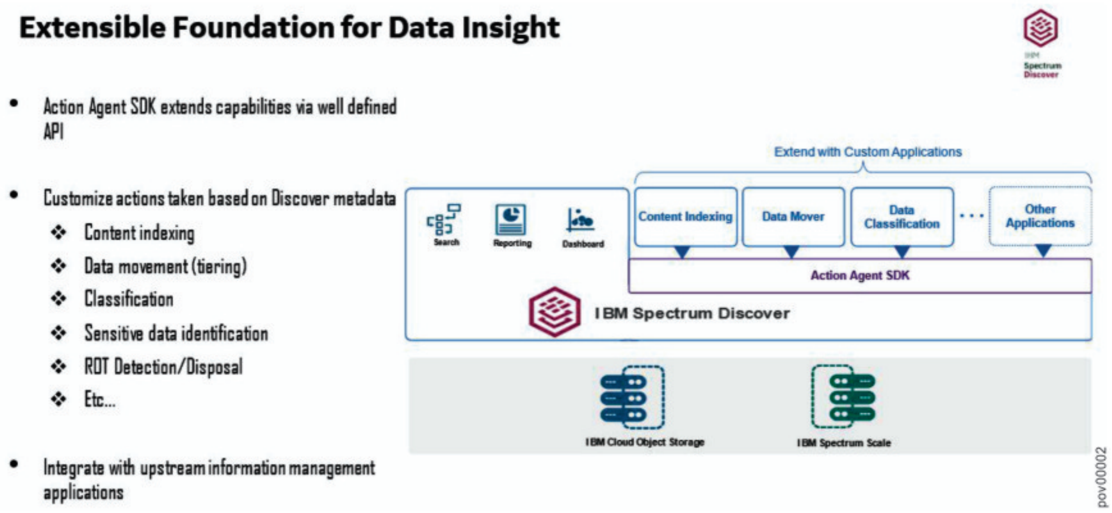


Figure 1. IBM Spectrum Discover architecture

The IBM Spectrum Discover Action Agent SDK allows users to customize actions taken based on the metadata collected by the platform, for example

- Content indexing
- Data movement (for example tiering)
- Sensitive data identification
- ROT detection and disposal
- Integration with upstream Information Management applications

Figure 2 on page 3 shows an example of the Action Agent SDK architecture.



Extensible Foundation for Data Insight

- Action Agent SDK extends capabilities via well defined API
- Customize actions taken based on Discover metadata
  - Content indexing
  - Data movement (tiering)
  - Classification
  - Sensitive data identification
  - ROT Detection/Disposal
  - Etc...
- Integrate with upstream information management applications

Figure 2. Action agent SDK architecture

Role-based access control

IBM Spectrum Discover provides access to resources based on roles. You can restrict access to information based on roles.

The role that is assigned to a user or group determines the privileges for that user or group. Users and groups can be associated with collections, which use policies that determine the metadata that is available to view.

User and group access can be authenticated by IBM Spectrum Discover, an LDAP server, or the IBM Cloud Object Storage System. The administrator can manage the user access functions.

Roles

Roles determine how users and groups access records or the IBM Spectrum Discover environment.

Remember: If a user or group is assigned to multiple roles, the least restrictive role is applicable.

For example, if you are assigned the role of **Data User**, and you are also assigned the role of a **Data Admin**, you have the privileges of a **Data Admin**.

Admin

An Admin can create users, groups, collections, manage LDAP, and IBM Cloud Object Storage connections for user access management.

Data Admin

Users with the **Data Admin** role can access all metadata that is collected by IBM Spectrum Discover and is not restricted by collections.

Collection Admin

The **Collection Admin** role is as a bridge between the **Data Admin** role and the **Data User** role. For example:

- Users with the **Collection Admin** role can list any type of tag and create or modify `Characteristic` tags. Users with the **Collection Admin** role cannot create, modify, or delete `Open` and `Restricted` tags. These permissions are the same permissions as the **Data User** role.

  Note: The built-in `Collection` tag is a special tag that can be set only by users with the **Data Admin** role. All other tags can be set by any user with the **Data User** or **Data Admin** or **Collection Admin** role.

- Users with the **Collection Admin** role can

- Create, update, and delete the policies for the collections they administer.
- View, update, and delete policies of data users for the collections they administer. They cannot delete a policy if it has a collection that they do not administer.
- Add users to collections that they administer. These data users have access to a particular collection, which means that they have access to the records marked with that collection value.

Important: The **Collection Admin** role is available as a technology preview in the 2.0.1.1 release. For limitations on the usage of the **Collection Admin** role, see the *IBM Spectrum Discover Release Notes*.

]

Data User

Users with the **Data User** role can access metadata that is collected by IBM Spectrum Discover. Metadata access might be restricted by policies in the collections that are assigned to users in this role. A user with the **Data User** role can also define tags and policies based on the collections to which the role is assigned.

Service User

The **Service User** role is assigned to accounts for IBM service and support personnel.

Data source connections

A data source connection specifies the parameters for cataloging of metadata from a source system to IBM Spectrum Discover.

Without the proper connection information, ingesting metadata from a connected system fails. You can use the **data source connections** page to view connection information for the data sources that are connected to your environment.

From the **data source connections** page, you can view the following information for the connections:

Source name

A name that uniquely identifies the connection to the data source. A data source can have multiple connections.

Platform

The domain of the data source - Spectrum Scale or IBM Cloud Object Storage System.

Cluster

The cluster address of the data source.

Data source name

The full name of the data source.

Site

The physical location of the data source.

See Table 3 on page 4 for an example of commands.

Table 3. API commands

API commands	
Command	Description
POST */connmgr/v1/connections*	• Creates new data source registration. • Takes the JSON input (with Name, Type, Cluster, dates source, Site) and adds an entry to the CONNECTIONS table.
GET */connmgr/v1/connections*	• Returns the contents for all the registered IBM Spectrum Discover Data Source connections.

Table 3. API commands (continued)

API commands	
GET */connmgr/v1/connections/<conn_name>*	• Returns the contents for all the registered IBM Spectrum Discover data source connections.
PUT */connmgr/v1/connections/<conn_name>*	• Updates the fields in the CONNECTIONS table where name == conn_name. • Only fields that are taken by POST are editable.
PUT */connmgr/v1/connections/<conn_name>*	• Updates the current_gen fields to the new value provided in the input form after all consumer partitions are processed. • Takes JSON input like *{"ten":, "quart":}*.
DELETE */connmgr/v1/connections/<conn_name>*	• Removes the row for name == conn_name and discard all information about the data source. • Need to have a confirmation to avoid accidental deletion. • Schedule implicit policy to delete all records associated with data source.

Cataloging metadata

Cataloging metadata in IBM Spectrum Discover is the process of ingesting and indexing the system metadata records from a source. Cataloging metadata transforms the metadata records into data that the user can act on and reference on.

Note: Metadata is data that describes data. Metadata captures the useful attributes of the associated source data to give the metadata context and meaning. For example, source data is a file or an object. The metadata is a set of attributes that are usually key: value pairs. The metadata records are associated with the file or object and are typically stored on the same system as the source data.

Note: System metadata is created and updated by the host system, and not the application software. IBM Spectrum Discover allows the addition of tags that can capture non-system metadata specific attributes.

The IBM Spectrum Discover data ingest pipeline uses Apache Kafka at its core to transfer the metadata records from the source system into the cluster.

IBM Spectrum Discover implements a set of Kafka producers and consumers for each source data type. The producers and consumers are within the cluster and handle the processing and normalization of the metadata records.

The source data types that are supported are as follows:

- IBM Spectrum Scale Scan.
- IBM Spectrum Scale Live Event.
- IBM Cloud Object Storage Scan.
- IBM Cloud Object Storage Live Event.

Scans are jobs that are scheduled or on demand, and occur at a data source level. For example, a file system or object vault. A set of metadata records is generated with each record that captures the state of an individual file or object within the data source at the time of the scan.

Live event notifications are triggered by user actions on the source data. Examples are reading, writing, moving, deleting data, changing permissions, or ownership. The Events generate a metadata record in real time that is stored in IBM Spectrum Discover.

Role of the producer

The IBM Spectrum Discover producer facilitates the process of reading source platform metadata records into the Apache Kafka cluster that sits in IBM Spectrum Discover. Each source data type has a dedicated producer to handle all incoming records. The records are provided to the producer through a single partition connector topic.

The producer partitions and publishes the records to a multi-partitioned target topic. The set of consumers at the other end of the target topic process the incoming records in parallel. Metadata records for a particular file or object are always routed to the same partition on the target topic. Using this method ensures that events for a file or object are always processed in the order they are received.

Role of the consumer

The IBM Spectrum Discover consumer's role is to normalize the system metadata records from the source system and store the results into the IBM Spectrum Discover database. Multiple consumers work in parallel, as a Kafka consumer group, for a single source data type. Each consumer in the group is assigned a single partition and processes the records from the partition in batches. Maintaining a set of consumers who work in parallel allows IBM Spectrum Discover a higher level of throughput for the record ingest.

Enriching metadata

IBM Spectrum Discover can enrich the metadata from supported platforms with additional information by using policies, action agents, and custom tags.

Tags

Custom tags are key:value pairs that are added to the IBM Spectrum Discover metadata record that allow the user to manage, report, and search for data by using newly applied information. The custom tags act as an extension of the system metadata that can contain organizational information beyond the view and limits of the source storage system. This information can be used for roll-up type aggregation and reporting or for targeted searching on specific values to find needles in haystacks.

Policies

Policies are used to add additional information about the source data that is indexed in IBM Spectrum Discover. A policy determines the set of files to add tag values to or send to an action agent through filtering criteria. The policies give the user the ability to run actions one time or on a set schedule. Policies do work in batches and can be paused, resumed, stopped, and restarted. The user can control the load on the IBM Spectrum Discover system and on the source storage system in the case of deep inspection policies.

Action agents

A deep inspect action agent does the work of extracting information from source data records and returning it to IBM Spectrum Discover to be indexed. For example, by using a custom action agent, a user might create a DEEPINSPECT policy to extract key characteristics from files of a certain type. The characteristics are applied to the metadata records for the files in IBM Spectrum Discover as custom tags and made searchable. A user can search for data by name, size, and content.

Tags

A tag is a custom metadata field that is used to supplement storage system metadata with organization-specific information.

An organization might segment their storage by project or by chargeback department. Those facets do not show in the system metadata and the storage systems themselves do not provide management and reporting capabilities based on those organizational concepts. With custom tags you can store additional information, and manage, report, and search for data by using that organizationally important information.

Types of tags

Categorization
Categorization tags contain values such as project, department, and security classification. Open and Restricted types of tags are Categorization tags. Size limit is 256 bytes.

Characteristic
Characteristic tags can contain any value that is needed to describe or classify the object. Can contain long descriptive values. Size limit is 4 KB.

Permissions

Security administrators
Cannot create, update, delete, or list any type of tag.

Data administrators
A data administrator can:

- Use READ and WRITE access to all tags.
- Create, modify, view, and delete all types of tags.

Data users
A data user can:

- Use READ only access to OPEN and RESTRICTED tags.
- Use READ and WRITE access to CHARACTERISTIC tags.
- View or list OPEN and RESTRICTED tags.
- Create or modify a CHARACTERISTIC tag.
- Create, modify, or delete OPEN or RESTRICTED tags.
- Not delete CHARACTERISTIC tags.

Policy engine

Policies offer a method whereby you can schedule one-time or repetitive actions on a filtered set of records.

The policy management API service is a RESTful web service that is designed to create, list, update, and delete policies. You can use a policy to initiate action on a select set of indexed documents or data. You can do a task immediately or on a set schedule.

Several types of policies that are supported by IBM Spectrum Discover enrich the metadata records. You can create policies with information to determine which set of documents to run, the action to take, and when to run policies periodically.

A policy includes

Policy ID
Name of the policy.

Filter
Selects a set of documents to work.

Action
Id, parameters, and schedule.

The following list is a description of the policies.

AUTOTAG
A policy that tags a set of records based on filter criteria with a pre-defined set of tags.

DEEPINSPECT

A policy that passes lists of files based on filter criteria to an analytics agent that opens the source data file and extracts metadata information from it. The policy passes the data back to IBM Spectrum Discover in the form of tags so you can do a search, and:

- Set up a filter to do a search query that finds the candidates to apply the policy.

  For example, you can set an action for filtered candidates AUTOTAG, tag1: value, tag2: value
- Set a schedule to apply the policy by specifying the following methods:
  - Immediately
  - Periodically

The following is an example of an AUTOTAG policy.

```
$ curl -k -H "Authorization: Bearer <token>"
https://<spectrum_discover_host>:443/policyengine/v1/policies/autotagpol1 -d '
{"pol_filter": "user='research1'", "action_id": "AUTOTAG", "action_params":
{"tags": {"tag1":"myTag1", "tag10":"proj1"}}}' -X POST -H "Content-Type: application/json"
```

The following is an example of a DEEPINSPECT policy.

```
{ "pol_id": "pol3", "action_id": "DEEPINSPECT", "action_params":
{ "agent": "myDeepInspect", "extract_tags": ["patient_name", "patient_age"] },
"schedule": "NOW", "pol_filter": "size>10000" }
```

Action agents

IBM Spectrum Discover policies might contain action agents in the actions parameters.

Use an action agent when you want to do a specific action on data or metadata on IBM Spectrum Discover.

You can define an agent when you create a new DEEPINSPECT policy. You can add parameters for an agent during the process of creating a DEEPINSPECT policy.

When you open the window for agents, you can see a view of a table with the following information:

Agent

The name of the agent.

Parameters

The parameters that were assigned to the agent when the policy was created.

Action ID

Deep inspect - the policy agent to which the agent is assigned.

View or Delete

Use the delete trashcan icon to remove the agent from the database.

Graphical user interface

The IBM Spectrum Discover graphical user interface is a portal that is used for running data searches, report generation, policy and tag management, and user Access Management. Based on a user's role, they might have access to one or more of these areas.

The IBM Spectrum Discover environment provides access to users and groups. The role that is assigned to a user or group determines the functions that are available. Users and groups can also be associated with collections, which use policies that determine the metadata that is available to view.

User and group access can be authenticated by IBM Spectrum Discover, an LDAP server, or the IBM Cloud Object Storage System. The administrator can manage the user access functions.

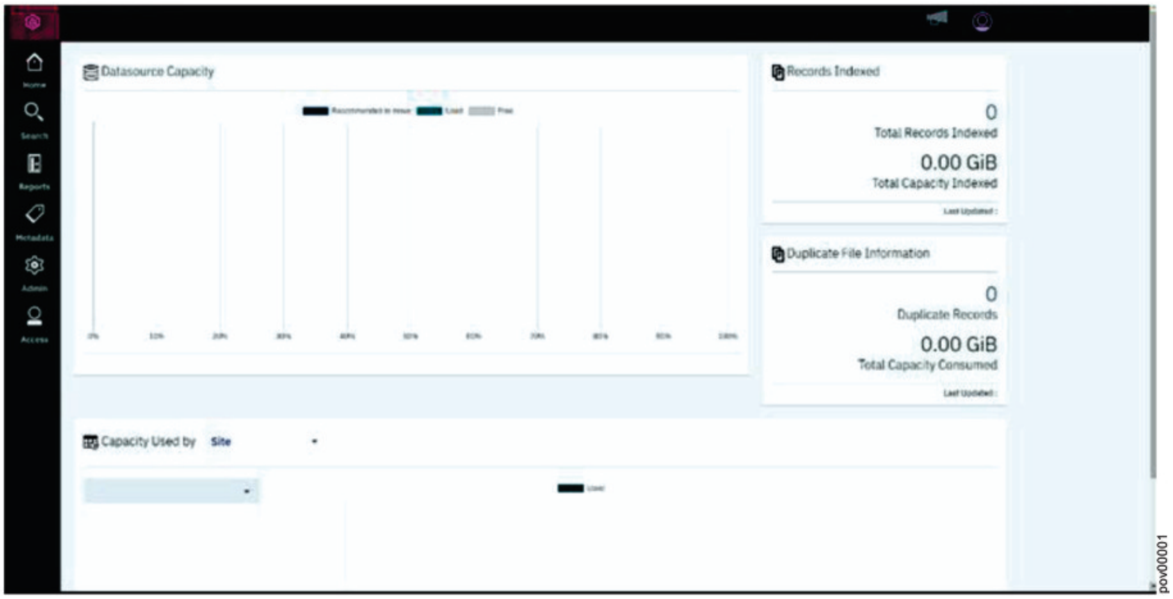
Roles

Roles determine how users and groups can access records or the IBM Spectrum Discover environment.

If a user or group is assigned to multiple roles, the least restrictive role is used. For example, if a user is assigned a role of Data User, and is included in a data administrator role, the user has the privileges of a data administrator.

Dashboard

Figure 3 on page 9 shows an example of an IBM Spectrum Discover dashboard.



Records Indexed

0
Total Records Indexed

0.00 GiB
Total Capacity Indexed

Last Updated :

Duplicate File Information

0
Duplicate Records

0.00 GiB
Total Capacity Consumed

Last Updated :

Figure 3. Example of the IBM Spectrum Discover dashboard

Data administrators and users can view the following:

- Metrics for the overall capacity used by every data source
- Total number of files
- Amount of capacity that is used by records with specific tags and facets, for example, owner, cluster, and size range
- Distribution of those records across data sources

Users can click any of the dashboard widgets to initiate a search and further explore and drill down into the data. Administrators and user can also perform the following:

- Monitor storage usage and data recommendations
- View total indexed data and capacity
- View duplicate file or object candidates. For example:
  - Number
  - Capacity used
- Preview capacity use by data facet - for example:
  - Classification
  - Owner
  - File type
- View data capacity by group or collection - for example:
  - Customer defined
  - Lab or project

Understanding size and capacity differences

[

IBM Spectrum Discover collects size and capacity information. Generally:

- Size refers to the size of a file or object in bytes.
- Capacity refers to the amount of space the file or object is consuming on the source storage in bytes.

For objects, size and capacity values always match. For files, size and capacity values can be different because of file system block overhead or sparsely populated files.

Note: Storage protection overhead (such as RAID values or erasure coding) and replication overhead are not captured in the capacity values.

]

Reports for IBM Spectrum Discover

Reports for IBM Spectrum Discover are grouped or non-grouped. Grouped reports have information for count and sum in columns and non-grouped reports have information in rows.

Data Curation Reports are a way for administrators, also known as data curators, to view the state of their storage environment in different ways. They can range from high-level grouped information to individual record level information.

For example, you can sort a report by owner, project, and department, or you can generate a list of records that meet a specific criteria. And, you can create a report that lists the records in a project that has not reviewed for over a year. The owner of the data can evaluate whether to archive or delete the report.

For more information on viewing and generating reports, see *Reports* in *IBM Spectrum Discover: Administration Guide*.

For information on using the IBM Spectrum Discover REST APIs to view and generate reports, see *IBM Spectrum Discover: REST API Guide*.

IBM Spectrum Discover appliance

The virtual appliance is a virtual machine in Open Virtualization Format (OVF) format that you can download and includes the IBM Spectrum Discover.

The IBM Spectrum Discover virtual appliance is bundled as Open Virtualization Appliance (OVA) image to be deployed on VMware vSphere 6.0 or later. vSphere is Vmware's hypervisor platform that is designed to manage large pools of virtualized computing infrastructure that includes software and hardware. Virtual appliance deployments use Vmware's ESXi hypervisor architecture.

The IBM Spectrum Discover virtual appliance cluster is automatically configured according to the input the user provides at the initial configuration console.

Each IBM Spectrum Discover virtual appliance is configured with the virtual resources.

Table 4. Virtual resources for the virtual appliance

Component	**Value**
RAM (GB)	64
CPU	16
For the ESX server, SCSI controller 0 is listed as LSI Logic. The second SCSI controller is listed as LSI Logic SAS.	VMware para virtual

Table 4. Virtual resources for the virtual appliance (continued)

Component	**Value**
Hard disk 1 **Note:** Three virtual disks (VMDK) are required including the disk that is created when installing the appliance from the OVA.	500 GB
Network adapter	VM network

Chapter 2. Planning

Software requirements

Virtual appliance specifications to use IBM Spectrum Discover at your site are as follows:

IBM Spectrum Discover is bundled as an Open Virtualization Appliance (OVA) image to be deployed on VMware vSphere 6.0 or later.

Table 5. Browser requirements for the IBM Spectrum Discover GUI

Browser	**Version**
Google Chrome	67 and higher
Firefox	60 ESR and higher ESR releases
Microsoft Edge	All versions

IBM Spectrum Discover deployment models

IBM Spectrum Discover can be deployed using a single node or multiple nodes.

IBM Spectrum Discover can be deployed in two modes depending on your business requirements. The modes are single node and multi-node deployments.

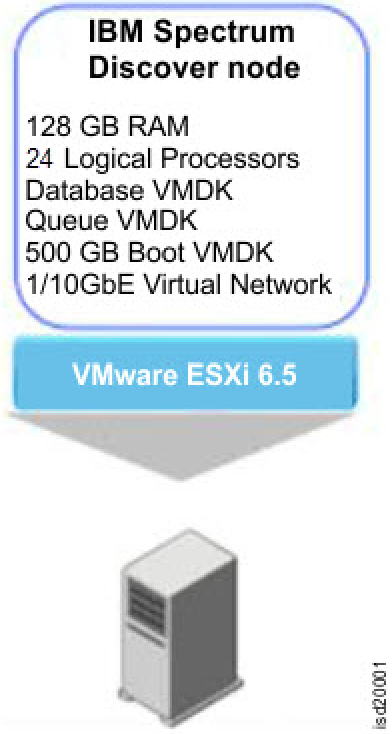


Figure 4. Single node deployment

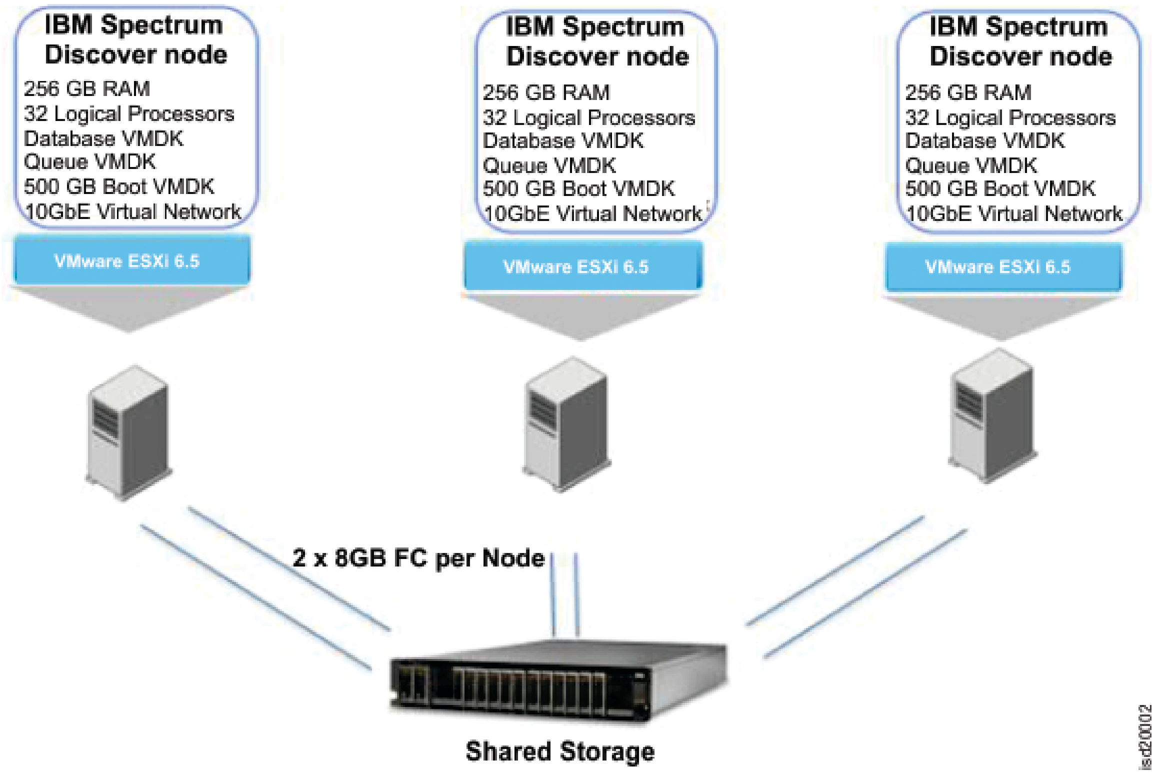


Figure 5. Multi-node deployment

CPU and memory requirements for single node trial and single node production IBM Spectrum Discover deployments

A description of the CPU and memory requirements for single node trial and single node production IBM Spectrum Discover deployment.

The following table shows the CPU and memory requirements for a single node production IBM Spectrum Discover deployment.

Table 6. CPU and memory requirements for single node production

Specification	**Value**
Memory	128 GB
Logical processors	24

The following table shows the recommended CPU and memory for a single node trial IBM Spectrum Discover deployment.

Note: Single node trial deployments with less than the recommended value of memory and logical processors will not be able to scale to index two billion documents.

Table 7. CPU and memory requirements for single node trial

Specification	**Minimum value**	**Recommended value**
Memory	64 GB	128 GB
Logical processors	8	24

Note: If using 64GB of RAM, no more than 25 million files can be indexed into IBM Spectrum Discover.

CPU and memory requirements for multi-node production IBM Spectrum Discover deployments

A description of the CPU and memory requirements for multi-node production IBM Spectrum Discover deployment.

The following table shows the CPU and memory requirements for a multi-node production IBM Spectrum Discover deployment.

Table 8. CPU and Memory Requirements for multi-node production

Specification	**Value**
Memory	256 GB
Logical processors	32

Networking requirements for IBM Spectrum Discover

IBM Spectrum Discover requires the following network parameters.

- Host name
- Virtual interface identifier
- IP address
- Netmask
- Gateway
- Domain Name Server (DNS) IP or host name
- Network Time Protocol (NTP) server IP or host name

Note: IBM Spectrum Discover requires a Fully Qualified Domain Name (FQDN) that is registered in a customer supplied DNS. The customer supplied FQDN must be resolvable by the customer supplied DNS from the IBM Spectrum Discover node in order for the IBM Spectrum Discover virtual appliance to operate properly.

The minimum recommended bandwidth for the network bandwidth is 1 GbE (Gigabit Ethernet) if action agent processing is not performed. If action agents are leveraged, the minimum recommended bandwidth is 10 GbE.

Note: The IBM Spectrum Discover nodes must be able to communicate with a customer supplied NTP server to operate properly.

Table 9. Network parameter example

Parameter	**Value Format**	**Recommended Value**	**Example**
Host name	`host.domain.com`	Fully qualified domain name (FQDN) of the node	`node1234.example.com`
Interface	`ensXXX`	The Ethernet interface to use for the virtual appliance networking	`ens192`
IP address	`xxx.xxx.xxx.xxx`	The IP address of the node	`10.10.200.10`
Netmask	`xxx.xxx.xxx.xxx`	Network mask for the IP range of the node	`255.255.255.0`

Table 9. Network parameter example (continued)

Parameter	Value Format	Recommended Value	Example
Gateway	`xxx.xxx.xxx.xxx`	IP address of the network gateway	`10.10.200.1`
DNS	`xxx.xxx.xxx.xxx`	The IP address of a single DNS	`10.10.200.35`
NTP	`xxx.xxx.xxx.xxx` or `host.domain.com`	Fully Qualified Domain Name or IP address of NTP server.	`10.10.10.2` or `Pool1.ntp.org`

Storage requirements for single node trial and single node production IBM Spectrum Discover deployments

This topic describes the storage requirements when you are using IBM Spectrum Discover as a single node trial deployment or a single node production deployment.

The single node IBM Spectrum Discover production appliance requires a 500 GB RAID protected SSD or flash Virtual Machine Disk) VMDK storage device for the operating system and base software. It is recommended that this VMDK be thick-provisioned and lazy-zeroed.

The single node production IBM Spectrum Discover virtual appliance requires an additional RAID protected SSD or flash VMDK storage device for the persistent message queue. The storage device for the persistent message queue can be locally attached storage or SAN-attached shared storage. It is recommended that this VMDK be thick-provisioned and lazy-zeroed. If an optional action agent is installed in the IBM Spectrum Discover node, additional storage capacity must be allocated for the VMDK storage device for the persistent message queue.

The single node production IBM Spectrum Discover virtual appliance also requires an additional RAID-protected SSD or flash Virtual Machine Disk (VMDK) storage device for the database.

You must add the persistent message queue and database VMDK storage devices to the IBM Spectrum Discover virtual appliance as part of the configuration process.

The following table shows the storage requirements for a single node production IBM Spectrum Discover deployment that supports indexing the metadata for up to 2 billion files and objects.

Table 10. Storage requirements for single node production

Use	Storage type	Size
Base OS and Software	Thick-provision and lazy-zero SSD / flash VMDK	500 GB
⟦Persistent message queue	Thick-provisioned and lazy-zero SSD / flash VMDK	700 GB⟧
Database (includes capacity for database backup)	Thick-provision and lazy-zero SSD / flash VMDK	2.5 TB

For single node non-production trial versions of IBM Spectrum Discover, a 500 GB RAID-protected HDD, SSD, or flash VMDK storage device is required for the operating system and base software. It is recommended that this VMDK be thick-provisioned and lazy-zeroed.

The single node non-production trial version of the IBM Spectrum Discover virtual appliance requires an additional RAID-protected HDD, SSD or flash Virtual Machine Disk (VMDK) storage device for the persistent message queue. If an optional action agent is installed in the IBM Spectrum Discover node, additional storage capacity must be allocated for the VMDK storage device for the persistent message queue.

An additional RAID-protected SSD or flash VMDK storage device is required for the database.

You must add the persistent message queue and database VMDK storage devices to the IBM Spectrum Discover virtual appliance as part of the configuration process. The two additional storage devices might be smaller in size than a single node production deployment if less than 2 B records will be indexed into the system. The following table shows the storage requirements.

Table 11. Storage requirements for single node trial

Use	**Storage type**	**Size**
Base OS and Software	Thick-provision and lazy-zero HDD or SSD / flash VMDK	500 GB
⟦Persistent message queue	Thick-provision and lazy-zero HDD or SSD / flash VMDK	50 GB minimum, 2 GB per 20 million indexed files⟧
Database (includes capacity for database backup)	Thick-provision and lazy-zero SSD / flash VMDK	100 GB minimum, 2 GB per 2 million indexed files
Database (does not include capacity for database backup)	Thick-provision and lazy-zero SSD / flash VMDK	100 GB minimum, 1 GB per 2 million indexed files

Storage requirements for multi-node production IBM Spectrum Discover deployments

A description of the storage requirements for multi-node IBM Spectrum Discover deployments.

A multi-node IBM Spectrum Discover deployment comprises three virtual appliance nodes.

Each node in the multi-node IBM Spectrum Discover production cluster requires a 500 GB RAID-protected SSD or flash VMDK storage device for the operating system and base software. It is recommended that this VMDK be thick-provisioned, lazy-zeroed. The storage device for the operating system and base software can be locally attached storage or SAN-attached shared storage.

Each node in the IBM Spectrum Discover virtual appliance node requires additional RAID-protected SSD or flash VMDK storage devices for the persistent message queue. The storage device for the persistent message queue can be locally-attached storage or SAN-attached shared storage. It is recommended that this VMDK be thick-provisioned and lazy-zeroed. If an optional action agent is installed in the IBM Spectrum Discover cluster, for each node, additional storage capacity must be allocated for the VMDK storage device for the persistent message queue.

Each node in the IBM Spectrum Discover virtual appliance node also requires a RAID-protected SSD or flash VMDK storage devices for the database. The storage device for the database must be SAN-attached shared storage. The VMDK for the database must be thick-provisioned and eager-zeroed.

Note: The database VMDK is shared between the IBM Spectrum Discover nodes in the IBM Spectrum Discover cluster. To share a VMDK between multiple nodes, VMware requires the volume to be thick-provisioned, eager-zeroed.

The persistent message queue and database storage devices are added to the IBM Spectrum Discover virtual appliance as part of the configuration process.

The following table shows the storage requirements for a three-node production IBM Spectrum Discover deployment that supports indexing the metadata for up to 10 billion files and objects:

Table 12. Storage requirements for multi-node production

Use	**Storage type**	**Size**
⟦Persistent message queue	Thick-provisioned and lazy-zero SSD / flash VMDK	1.4 TB⟧

Table 12. Storage requirements for multi-node production (continued)

Use	Storage type	Size
Database (includes capacity for database backup)	Thick provision, eager zero SSD / flash VMDK	14 TB

For multi-node deployments containing more than 10 billion files and objects, 2 GB per 2 million indexed files is required.

IBM Spectrum Scale and IBM Cloud Object Storage source software requirements

IBM Spectrum Discover indexes metadata from IBM Cloud Object Storage (COS) by receiving notifications containing metadata from COS and also supports scanning COS to harvest metadata.

The following table shows the minimum required COS software version to enable metadata harvesting with IBM Spectrum Discover:

Table 13. IBM Cloud Object Storage software requirements

Component	Version
IBM Cloud Object Storage (COS)	3.14.0 and higher

IBM Spectrum Discover indexes metadata from IBM Spectrum Scale by scanning IBM Spectrum Scale file systems. The IBM Spectrum Scale watch folders technical preview also enables IBM Spectrum Scale to send events containing metadata to IBM Spectrum Discover.

The following table lists the minimum required IBM Spectrum Scale software versions to enable metadata harvesting with IBM Spectrum Discover:

Table 14. IBM Spectrum Scale software requirements

Component	Metadata harvest method	Version
IBM Spectrum Scale	Scanning	4.2.3.x and higher
IBM Spectrum Scale	Live events technical preview	5.0.2.1 and higher

Backup and restore storage requirements for IBM Spectrum Discover

IBM Spectrum Discover provides a set of scripts for safely backing up and restoring the metadata database and file system.

The script integrates with the following backup targets:

- IBM Cloud Object Storage
- IBM Spectrum Protect
- External FTP server

The size of the backup pool for the backup targets is determined by taking the size of the backup staging pool x the number of backups kept as part of the retention policy.

Example:

Single node backup staging pool = 2 TB

Number of backups = 7

Backup target capacity required = 2 TB x 7 = 14 TB

Single node IBM Spectrum Discover production deployment planning worksheet

Use this worksheet to plan for installing IBM Spectrum Discover for a single node production deployment.

Note: All IBM Spectrum Discover Deployment Planning Worksheets can be downloaded here: IBM_Spectrum_Discover_Deployment_Planning_Worksheets.pdf

Table 15. Single node IBM Spectrum Discover production deployment planning

<table>
<tr><th colspan="5">CPU and memory requirements</th></tr>
<tr><th colspan="2">Parameter</th><th colspan="2">Recommended value</th><th>Record your values</th></tr>
<tr><td colspan="2">Memory</td><td colspan="2">128 GB</td><td></td></tr>
<tr><td colspan="2">Logical processor count</td><td colspan="2">24 logical processors</td><td></td></tr>
<tr><th colspan="5">Networking requirements</th></tr>
<tr><th>Parameter</th><th>Value format</th><th>Recommended value</th><th>Example</th><th>Record your values</th></tr>
<tr><td><hostname></td><td>host.domain.com</td><td>Fully qualified domain name of the node</td><td>node.example.com</td><td></td></tr>
<tr><td><interface></td><td>ensXXX</td><td>The Ethernet interface to use for the virtual appliance networking</td><td>ens192</td><td></td></tr>
<tr><td><ip></td><td>xxx.xxx.xxx.xxx</td><td>The IP address of the node</td><td>10.10.200.10</td><td></td></tr>
<tr><td><netmask></td><td>xxx.xxx.xxx.xxx</td><td>Network mask for the IP range of the node</td><td>255.255.254.0</td><td></td></tr>
<tr><td><gateway></td><td>xxx.xxx.xxx.xxx</td><td>IP address of the network gateway</td><td>10.10.200.1</td><td></td></tr>
<tr><td><dns></td><td>xxx.xxx.xxx.xxx</td><td>The IP address of a single DNS server</td><td>10.10.200.35</td><td></td></tr>
<tr><td><ntp></td><td>xxx.xxx.xxx.xxx or
host.domain.com</td><td>Fully Qualified Domain Name or IP address of NTP server.</td><td>Pool1.ntp.org</td><td></td></tr>
<tr><th colspan="5">Storage requirements</th></tr>
<tr><th colspan="2">Parameter</th><th colspan="2">Recommended value</th><th>Record your values</th></tr>
<tr><td colspan="2">Base OS SW VMDK</td><td colspan="2">500 GB thick provision, lazy zero SSD / flash</td><td></td></tr>
</table>

Table 15. Single node IBM Spectrum Discover production deployment planning (continued)

Persistent message queue VMDK	Persistent message queue: 700GB thick-provision, lazy-zero SSD / flash VMDK	
	Database VMDK	2.5 TB thick provision, lazy zero SSD / flash

Single node IBM Spectrum Discover trial deployment planning worksheet

Use this worksheet to plan for installing IBM Spectrum Discover for a single node trial deployment.

Note: All IBM Spectrum Discover Deployment Planning Worksheets can be downloaded here: IBM Spectrum Discover Deployment Planning Worksheets.pdf

Table 16. Single node IBM Spectrum Discover trial deployment planning

CPU and memory requirements

Parameter	Recommended value	Record your values
Memory	64 GB minimum 128 GB recommended	
Logical processor count	8 logical processors minimum 24 logical processors recommended	

Networking requirements

Parameter	Value format	Recommended value	Example	Record your values
<hostname>	`host.domain.com`	Fully qualified domain name of the node	`node.example.com`	
<interface>	ensXXX	The Ethernet interface to use for the virtual appliance networking	ens192	
<ip>	`xxx.xxx.xxx.xxx`	The IP address of the node	`10.10.200.10`	
<netmask>	`xxx.xxx.xxx.xxx`	Network mask for the IP range of the node	`255.255.254.0`	
<gateway>	`xxx.xxx.xxx.xxx`	IP address of the network gateway	`10.10.200.1`	
<dns>	`xxx.xxx.xxx.xxx`	The IP address of a single DNS server	`10.10.200.35`	
<ntp>	`xxx.xxx.xxx.xxx` or `host.domain.com`	Fully Qualified Domain Name or IP address of NTP server.	`Pool1.ntp.org`	

Table 16. Single node IBM Spectrum Discover trial deployment planning (continued)

<table>
<tr><th colspan="4">Storage requirements</th></tr>
<tr><th>Parameter</th><th colspan="2">Recommended value</th><th>Record your values</th></tr>
<tr><td>Base OS SW VMDK</td><td colspan="2">500 GB thick provision, lazy zero SSD / flash</td><td></td></tr>
<tr><td rowspan="3">Persistent message queue VMDK</td><td colspan="2">Persistent message queue: 50 GB minimum + 2 GB per 20 million indexed files, thick-provision, lazy-zero HDD or SSD / flash</td><td></td></tr>
<tr><td rowspan="2">Database VMDK</td><td colspan="2">Database (does not include capacity for database backup): 100 GB minimum, 1 GB per 2 million indexed files, thick provision, lazy zero SSD/flash VMDK</td></tr>
<tr><td>Database (includes capacity for database backup): 100 GB minimum, 2 GB per 2 million indexed files, thick provision, lazy zero SSD/flash VMDK</td><td></td></tr>
</table>

Note: If using 64GB of RAM, no more than 25 million files can be indexed into IBM Spectrum Discover.

Multi-node IBM Spectrum Discover production deployment planning worksheets

Use these planning worksheets to prepare for installing IBM Spectrum Discover for a multi-node production deployment for 10 billion indexed documents.

Note: All IBM Spectrum Discover Deployment Planning Worksheets can be downloaded here: IBM Spectrum Discover Deployment Planning Worksheets.pdf

Node 1

Table 17. Node 1: IBM Spectrum Discover production deployment planning

CPU and memory requirements

Parameter	Recommended value	Record your values
Memory	256 GB	
Logical processor count	32 logical processors	

Networking requirements

Parameter	Value format	Recommended value	Example	Record your values
<hostname>	`host.domain.com`	Fully qualified domain name of the node	`node.example.com`	

Table 17. Node 1: IBM Spectrum Discover production deployment planning (continued)

<interface>	ensXXX	The Ethernet interface to use for the virtual appliance networking	ens192	
<ip>	xxx.xxx.xxx.xxx	The IP address of the node	10.10.200.10	
<netmask>	xxx.xxx.xxx.xxx	Network mask for the IP range of the node	255.255.254.0	
<gateway>	xxx.xxx.xxx.xxx	IP address of the network gateway	10.10.200.1	
<dns>	xxx.xxx.xxx.xxx	The IP address of a single DNS server	10.10.200.35	
<ntp>	xxx.xxx.xxx.xxx or host.domain.com	Fully Qualified Domain Name or IP address of NTP server.	Pool1.ntp.org	
Storage requirements				
Parameter		**Recommended value**		**Record your values**
Base OS SW VMDK		500GB thick provision, lazy zero SSD / flash		
Persistent message queue VMDK		Persistent message queue: 1.4 TB, thick-provision, lazy-zero SSD / flash		
		Database VMDK	14 TB thick provision, eager zero SSD / flash. Shared VMDK between node1, node2, node3.	

Node 2

Table 18. Node 2: IBM Spectrum Discover production deployment planning.

Use this worksheet to plan for node 2 of an IBM Spectrum Discover production deployment for 10 billion indexed documents.

CPU and memory requirements				
Parameter		**Recommended value**		**Record your values**
Memory		256 GB		
Logical processor count		32 logical processors		
Networking requirements				
Parameter	**Value format**	**Recommended value**	**Example**	**Record your values**
<hostname>	host.domain.com	Fully qualified domain name of the node	node.example.com	

Table 18. Node 2: IBM Spectrum Discover production deployment planning.

Use this worksheet to plan for node 2 of an IBM Spectrum Discover production deployment for 10 billion indexed documents.

(continued)

<interface>	ensXXX	The Ethernet interface to use for the virtual appliance networking	ens192	
<ip>	xxx.xxx.xxx.xxx	The IP address of the node	10.10.200.10	
<netmask>	xxx.xxx.xxx.xxx	Network mask for the IP range of the node	255.255.254.0	
<gateway>	xxx.xxx.xxx.xxx	IP address of the network gateway	10.10.200.1	
<dns>	xxx.xxx.xxx.xxx	The IP address of a single DNS server	10.10.200.35	
<ntp>	xxx.xxx.xxx.xxx or host.domain.com	Fully Qualified Domain Name or IP address of NTP server.	Pool1.ntp.org	

Storage requirements

Parameter	Recommended value	Record your values
Base OS SW VMDK	500 GB thick provision, lazy zero SSD / flash	
Persistent message queue VMDK	Persistent message queue : 1.4 TB thick provision, lazy zero SSD / flash VMDK	
	Database VMDK	14 TB thick provision, eager zero SSD / flash. Shared VMDK between node1, node2, node3.

Node 3

Table 19. Node 3: IBM Spectrum Discover production deployment planning.

Use this worksheet to plan for node 3 of an IBM Spectrum Discover production deployment for 10 billion indexed documents.

CPU and memory requirements

Parameter	Recommended value	Record your values
Memory	256 GB	
Logical processor count	32 logical processors	

Networking requirements

Parameter	Value format	Recommended value	Example	Record your values

Table 19. Node 3: IBM Spectrum Discover production deployment planning.

Use this worksheet to plan for node 3 of an IBM Spectrum Discover production deployment for 10 billion indexed documents.

(continued)

<hostname>	`host.domain.com`	Fully qualified domain name of the node	`node.example.com`	
<interface>	`ensXXX`	The Ethernet interface to use for the virtual appliance networking	`ens192`	
<ip>	`xxx.xxx.xxx.xxx`	The IP address of the node	`10.10.200.10`	
<netmask>	`xxx.xxx.xxx.xxx`	Network mask for the IP range of the node	`255.255.254.0`	
<gateway>	`xxx.xxx.xxx.xxx`	IP address of the network gateway	`10.10.200.1`	
<dns>	`xxx.xxx.xxx.xxx`	The IP address of a single DNS server	`10.10.200.35`	
<ntp>	`xxx.xxx.xxx.xxx` or `host.domain.com`	Fully Qualified Domain Name or IP address of NTP server.	`Pool1.ntp.org`	
Storage requirements				
Parameter		**Recommended value**		**Record your values**
Base OS SW VMDK		500 GB thick provision, lazy zero SSD / flash		
Persistent message queue VMDK		Persistent message queue: 1.4 TB thick provision, lazy zero SSD / flash		
		Database VMDK	14 TB thick provision, eager zero SSD / flash. Shared VMDK between node1, node2, node3.	

Chapter 3. Deploying and configuring

This section provides information on how to deploy and configure IBM Spectrum Discover single node trial, single node, or multi-node production virtual appliance.

Deploy and configure a single node production IBM Spectrum Discover appliance cluster

This section provides information on how to deploy and configure IBM Spectrum Discover single node trial or single node production virtual appliance.

Deploying a single node trial or single node production IBM Spectrum Discover virtual appliance

The IBM Spectrum Discover software is available as an OVA (open virtualization appliance) file. You can deploy it on your VMware ESXi server by using the **VMware vSphere Client**.

Before you begin

- Download the IBM Spectrum Discover OVA file on the local system or obtain the URL to an IBM Spectrum Discover OVA file accessible on the internet.
- Review deployment and configuration known issues and workarounds. For more information, see "Known issues with deploying and configuring for single node" on page 45.

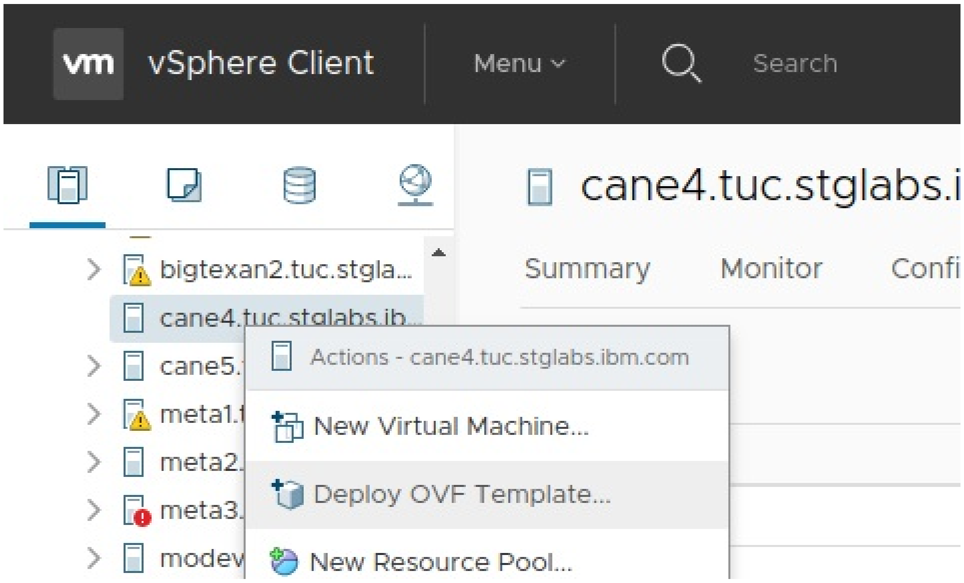
Important: You cannot use the HTML5 vSphere client to deploy the OVF image.

About this task

Deploy the IBM Spectrum Discover virtual appliance as follows by using the **Deploy OVF Template** wizard of the VMware vSphere Client.

Procedure

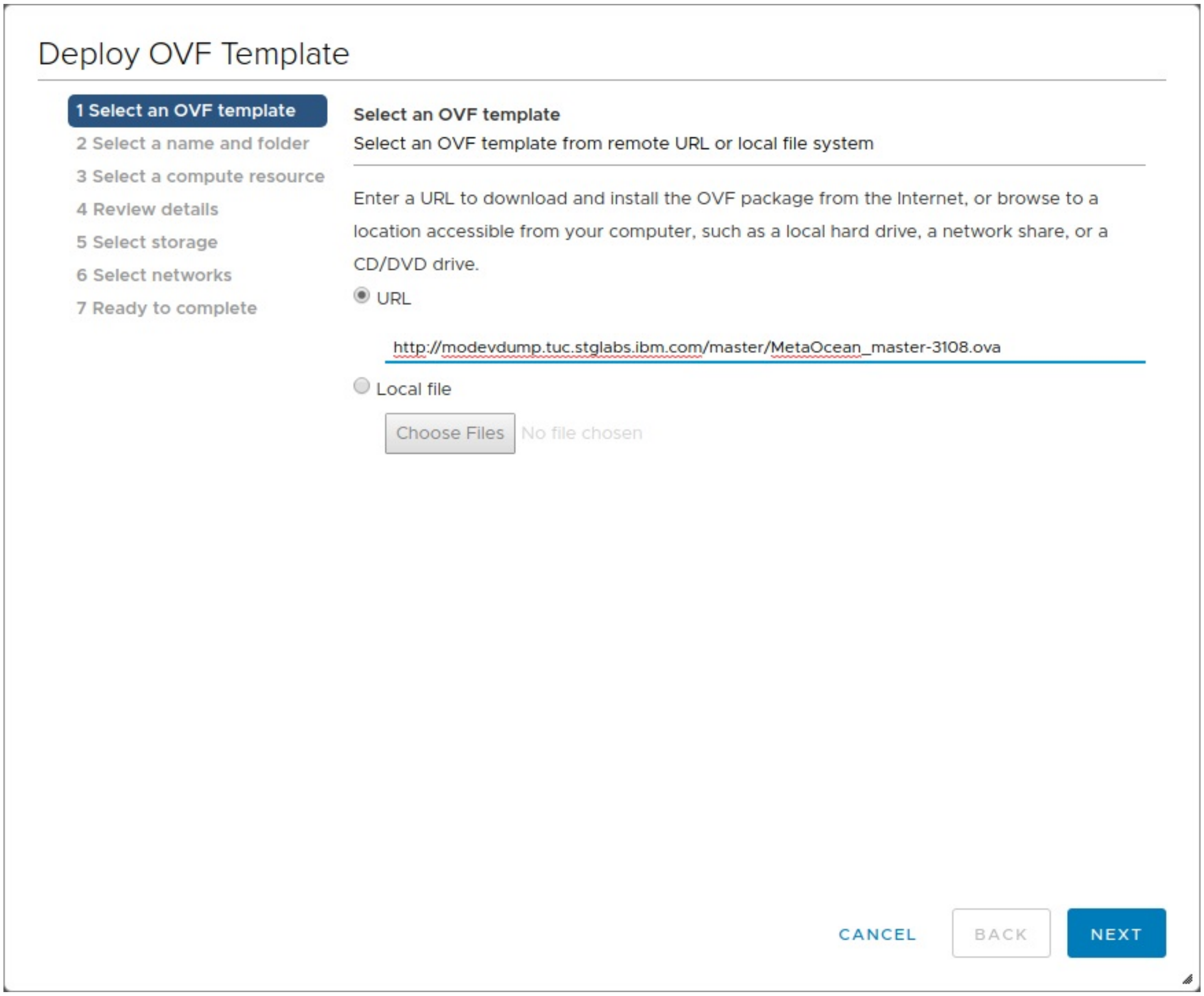
1. In the vSphere Client, right-click the ESXi server on which you want to deploy the virtual appliance and then click **Deploy OVF Template**.



The **Deploy OVF Template** wizard appears.

2. Select the IBM Spectrum Discover virtual appliance that you want to deploy and click **Next**.

   You can either select an OVA file that you have downloaded on the local system or you can specify a URL to the OVA file.



Deploy OVF Template

1 Select an OVF template
2 Select a name and folder
3 Select a compute resource
4 Review details
5 Select storage
6 Select networks
7 Ready to complete

Select an OVF template

Select an OVF template from remote URL or local file system

Enter a URL to download and install the OVF package from the Internet, or browse to a location accessible from your computer, such as a local hard drive, a network share, or a CD/DVD drive.

● URL

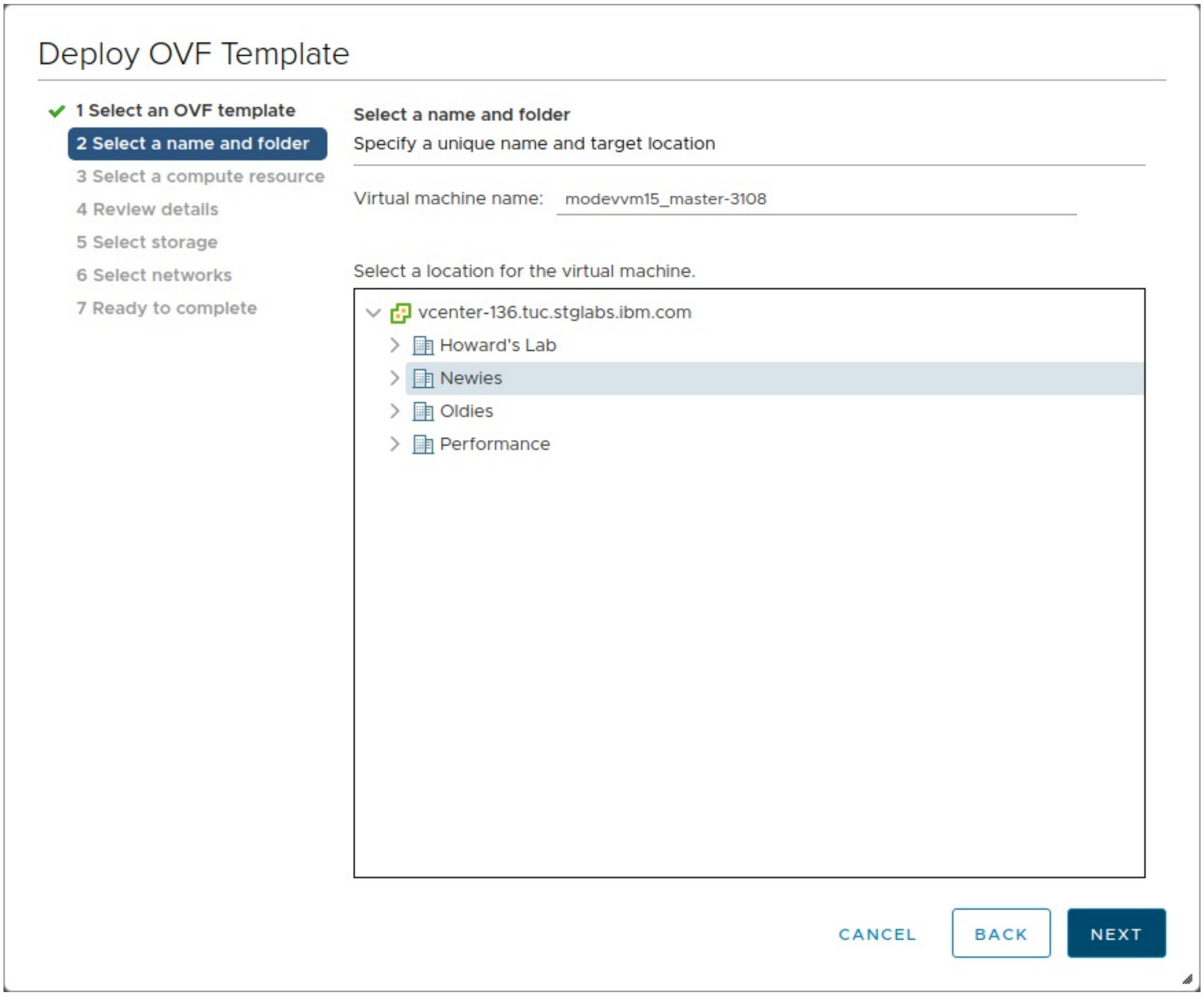
http://modevdump.tuc.stglabs.ibm.com/master/MetaOcean_master-3108.ova

○ Local file

Choose Files No file chosen

CANCEL BACK NEXT

3. Specify the name of the virtual appliance or accept the default name and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template

2 Select a name and folder

3 Select a compute resource

4 Review details

5 Select storage

6 Select networks

7 Ready to complete

Select a name and folder

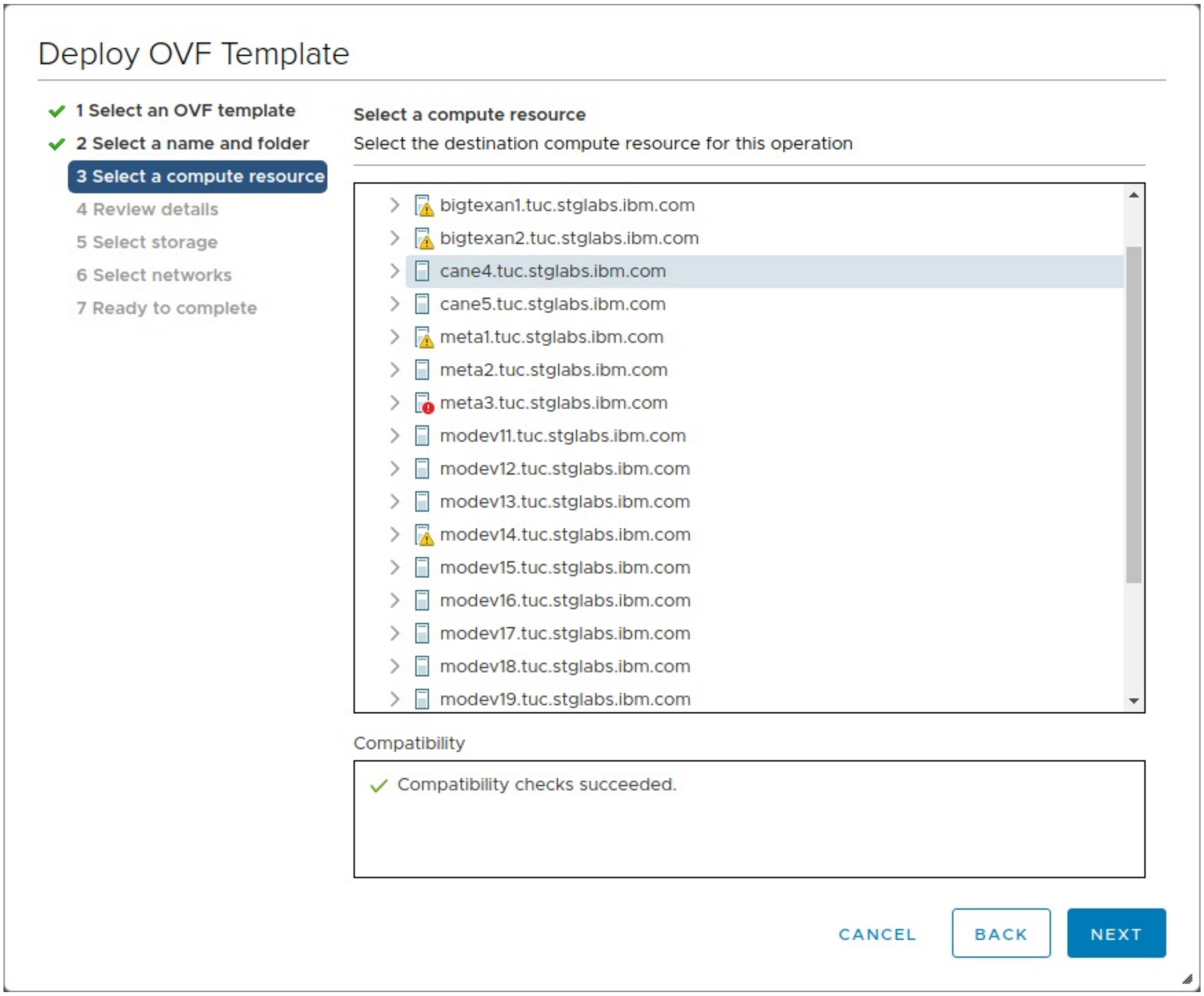
Specify a unique name and target location

Virtual machine name: modevvm15_master-3108

Select a location for the virtual machine.

- vcenter-136.tuc.stglabs.ibm.com
  - Howard's Lab
  - Newies
  - Oldies
  - Performance

4. Select the physical server on which you want to deploy the virtual appliance and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template
✓ 2 Select a name and folder
3 Select a compute resource
4 Review details
5 Select storage
6 Select networks
7 Ready to complete

Select a compute resource

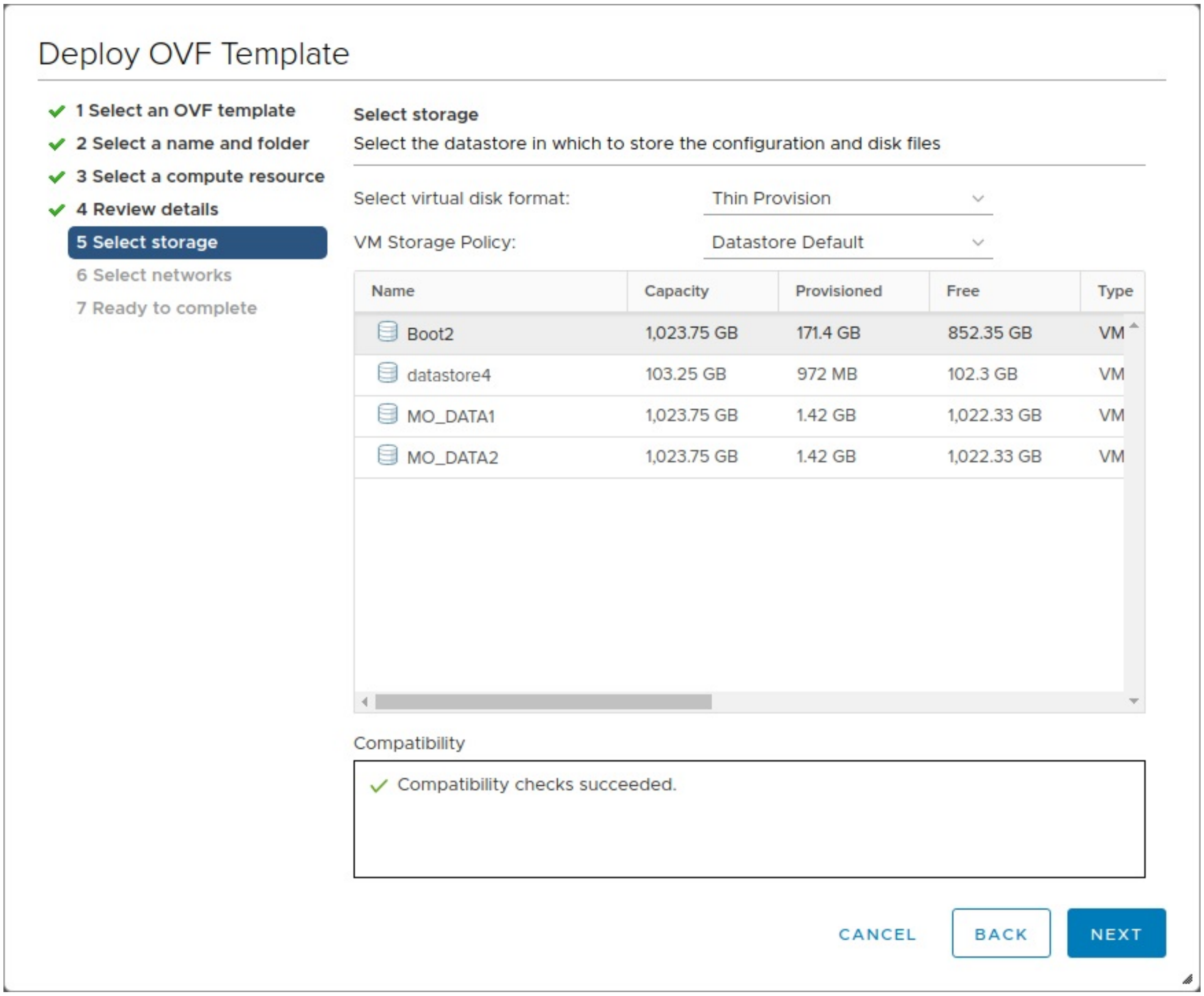
Select the destination compute resource for this operation

- bigtexan1.tuc.stglabs.ibm.com
- bigtexan2.tuc.stglabs.ibm.com
- cane4.tuc.stglabs.ibm.com
- cane5.tuc.stglabs.ibm.com
- meta1.tuc.stglabs.ibm.com
- meta2.tuc.stglabs.ibm.com
- meta3.tuc.stglabs.ibm.com
- modev11.tuc.stglabs.ibm.com
- modev12.tuc.stglabs.ibm.com
- modev13.tuc.stglabs.ibm.com
- modev14.tuc.stglabs.ibm.com
- modev15.tuc.stglabs.ibm.com
- modev16.tuc.stglabs.ibm.com
- modev17.tuc.stglabs.ibm.com
- modev18.tuc.stglabs.ibm.com
- modev19.tuc.stglabs.ibm.com

Compatibility

✓ Compatibility checks succeeded.

5. Review the details and click **Next**.
6. Select the check box to accept the terms of the licenses and click **Next**.
7. Select the data store for the virtual appliance and the virtual disk format, and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template
✓ 2 Select a name and folder
✓ 3 Select a compute resource
✓ 4 Review details
5 Select storage
6 Select networks
7 Ready to complete

Select storage

Select the datastore in which to store the configuration and disk files

Select virtual disk format: Thin Provision

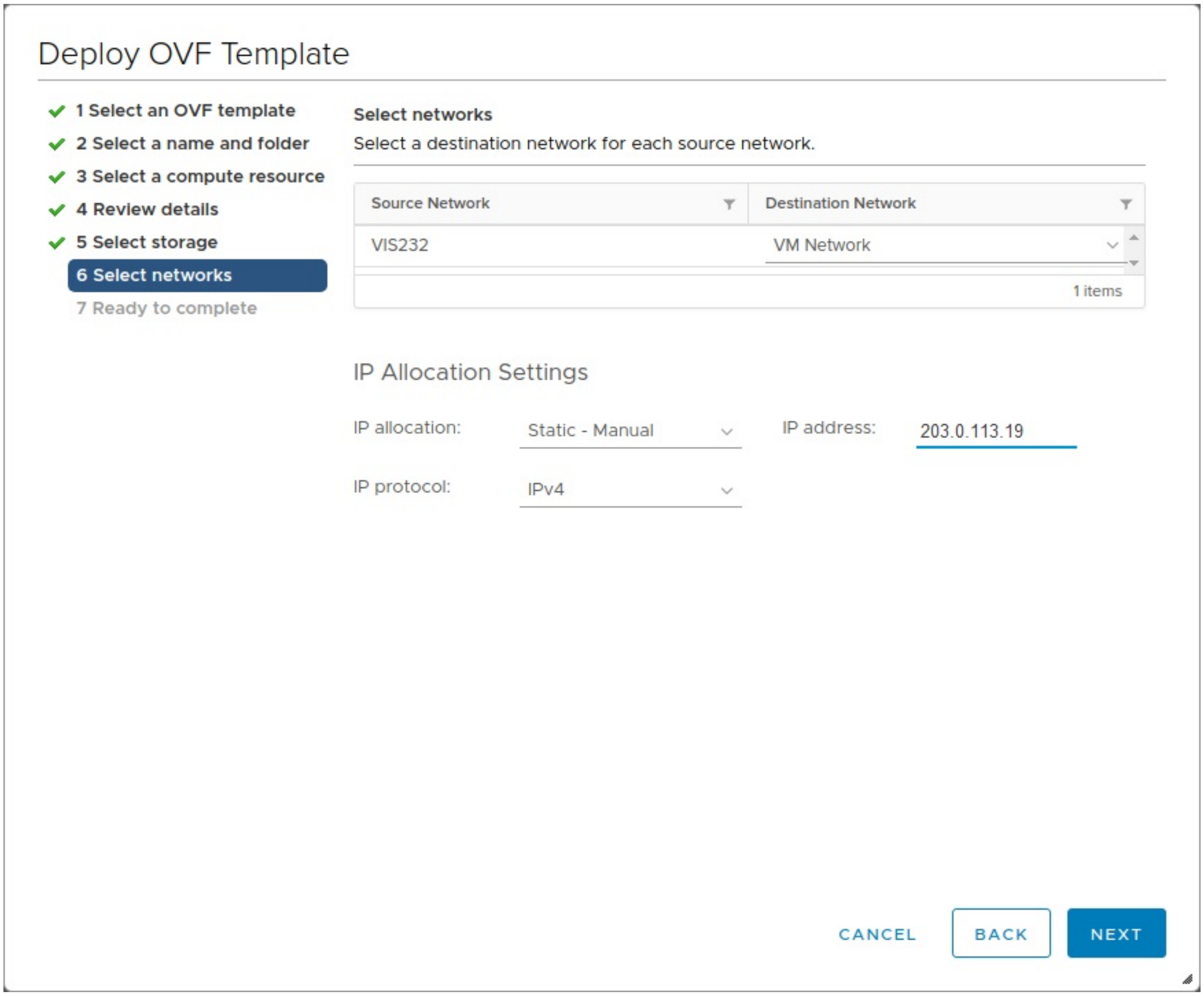
VM Storage Policy: Datastore Default

Name	Capacity	Provisioned	Free	Type
Boot2	1,023.75 GB	171.4 GB	852.35 GB	VM
datastore4	103.25 GB	972 MB	102.3 GB	VM
MO_DATA1	1,023.75 GB	1.42 GB	1,022.33 GB	VM
MO_DATA2	1,023.75 GB	1.42 GB	1,022.33 GB	VM

Compatibility

✓ Compatibility checks succeeded.

8. Select the VM network for the virtual appliance and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template
✓ 2 Select a name and folder
✓ 3 Select a compute resource
✓ 4 Review details
✓ 5 Select storage
6 Select networks
7 Ready to complete

Select networks

Select a destination network for each source network.

Source Network	Destination Network
VIS232	VM Network

1 items

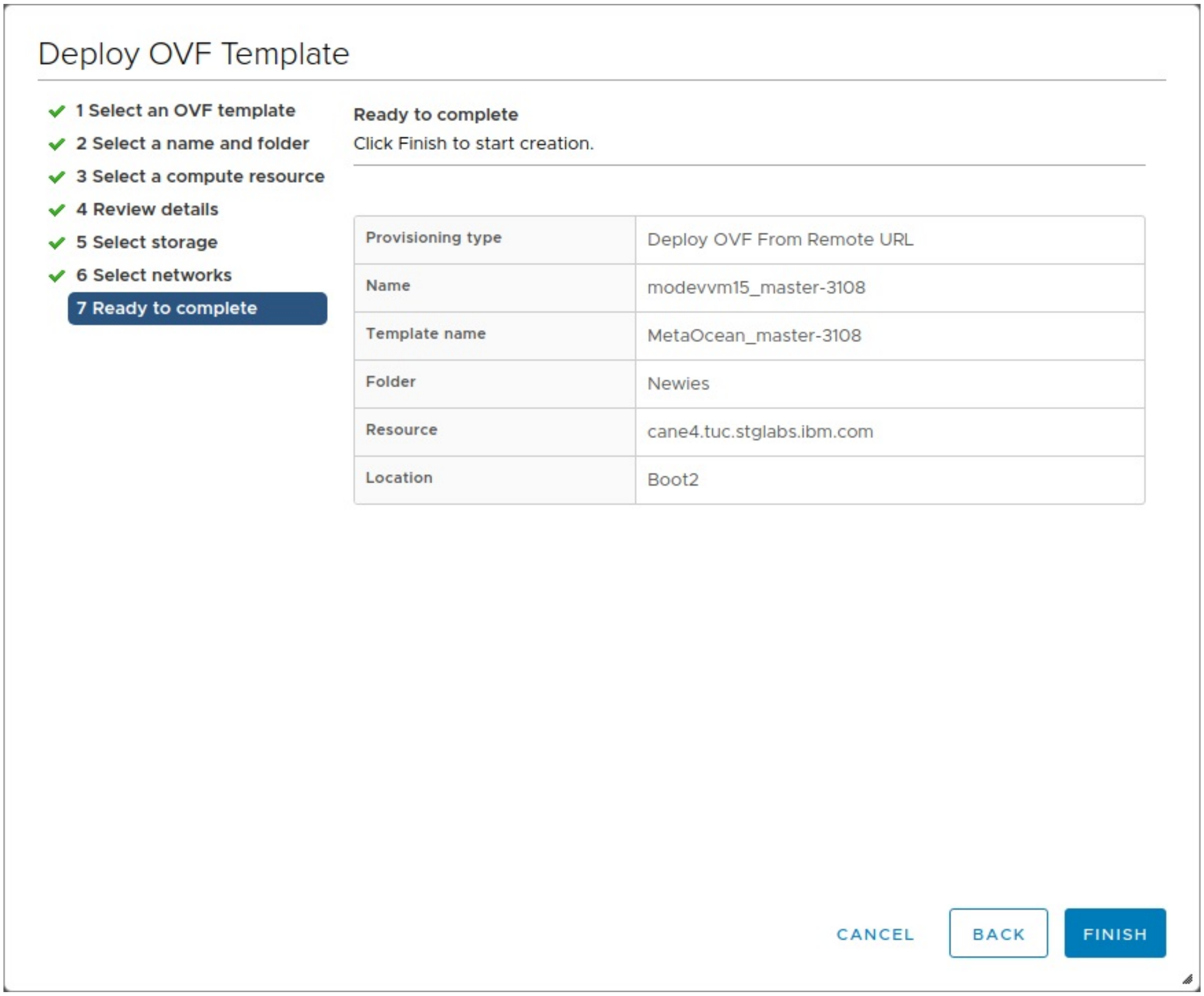
IP Allocation Settings

IP allocation: Static - Manual

IP address: 203.0.113.19

IP protocol: IPv4

9. Review the settings and click **Finish**.



Ready to complete

Click Finish to start creation.

Provisioning type	Deploy OVF From Remote URL
Name	modevvm15_master-3108
Template name	MetaOcean_master-3108
Folder	Newies
Resource	cane4.tuc.stglabs.ibm.com
Location	Boot2

The IBM Spectrum Discover virtual node gets created and the storage is provisioned.

Note: Do not power on the virtual appliance until storage, CPU, and memory have been configured.

Configuring storage for a single node trial or single node production of IBM Spectrum Discover virtual appliance

The IBM Spectrum Discover trial and production virtual appliance node requires two additional VMDK storage devices.

Note: The persistent message queue and database VMDK storage devices are in addition to the base OS and software VMDK that was automatically configured during the initial IBM Spectrum Discover virtual appliance deployment. A total of three virtual disks (VMDK) are required including the disk that is created when installing the appliance on the OVA.

Procedure

1. Add virtual disk for the IBM Spectrum Discover persistent message queues.

   For more information, see "Adding virtual disk for IBM Spectrum Discover persistent message queues" on page 32.
2. Add virtual disk for the IBM Spectrum Discover database.

   For more information, see "Adding a virtual disk for the IBM Spectrum Discover database" on page 35.

Adding virtual disk for IBM Spectrum Discover persistent message queues

You can use the VMware vSphere Client to add the virtual disk required for IBM Spectrum Discover persistent message queues to the virtual appliance.

About this task

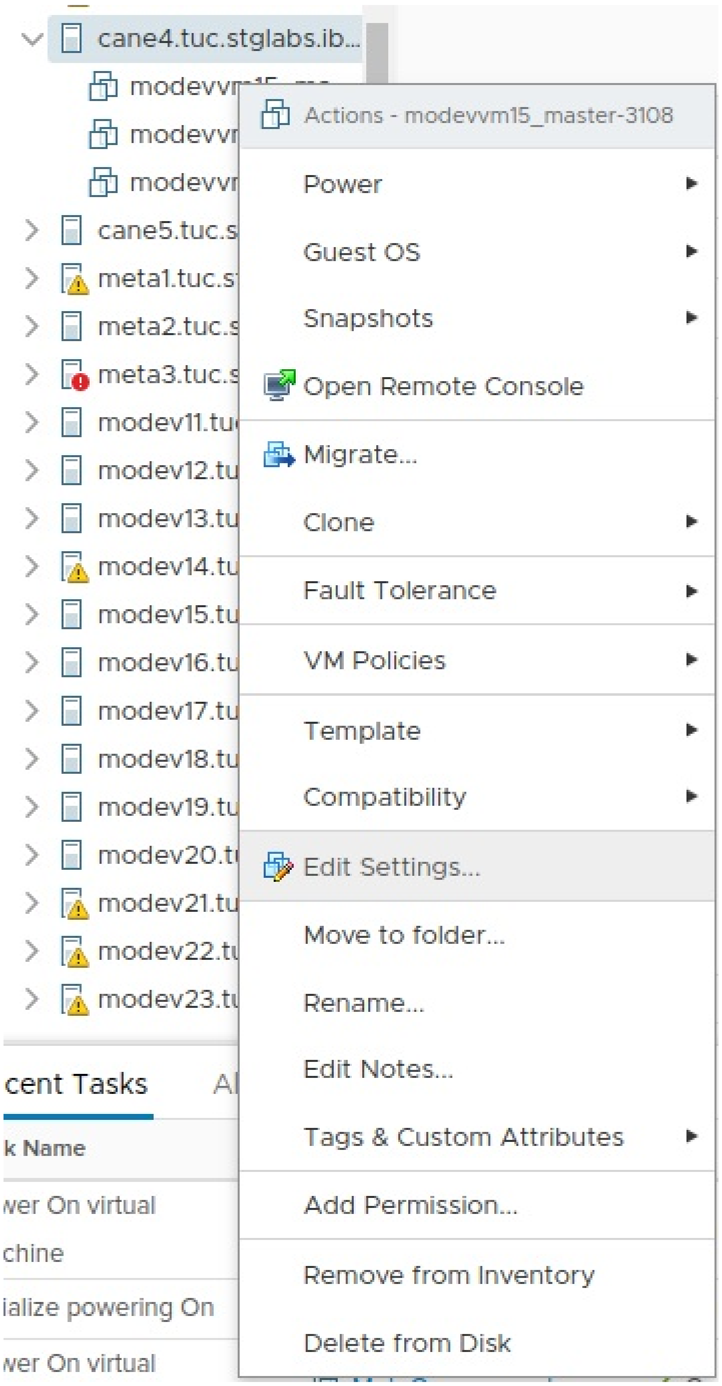
Important:

See the Planning section for detailed requirements for the persistent message queue VMDK. For a single node production IBM Spectrum Discover deployment, a 4.5 TB thick provisioned, and lazy zeroed VMDK is required. See the Planning section in the IBM Spectrum Discover Knowledge Center. If an optional IBM Spectrum Discover action agent is to be configured, an additional 1.6 TB of capacity is required.

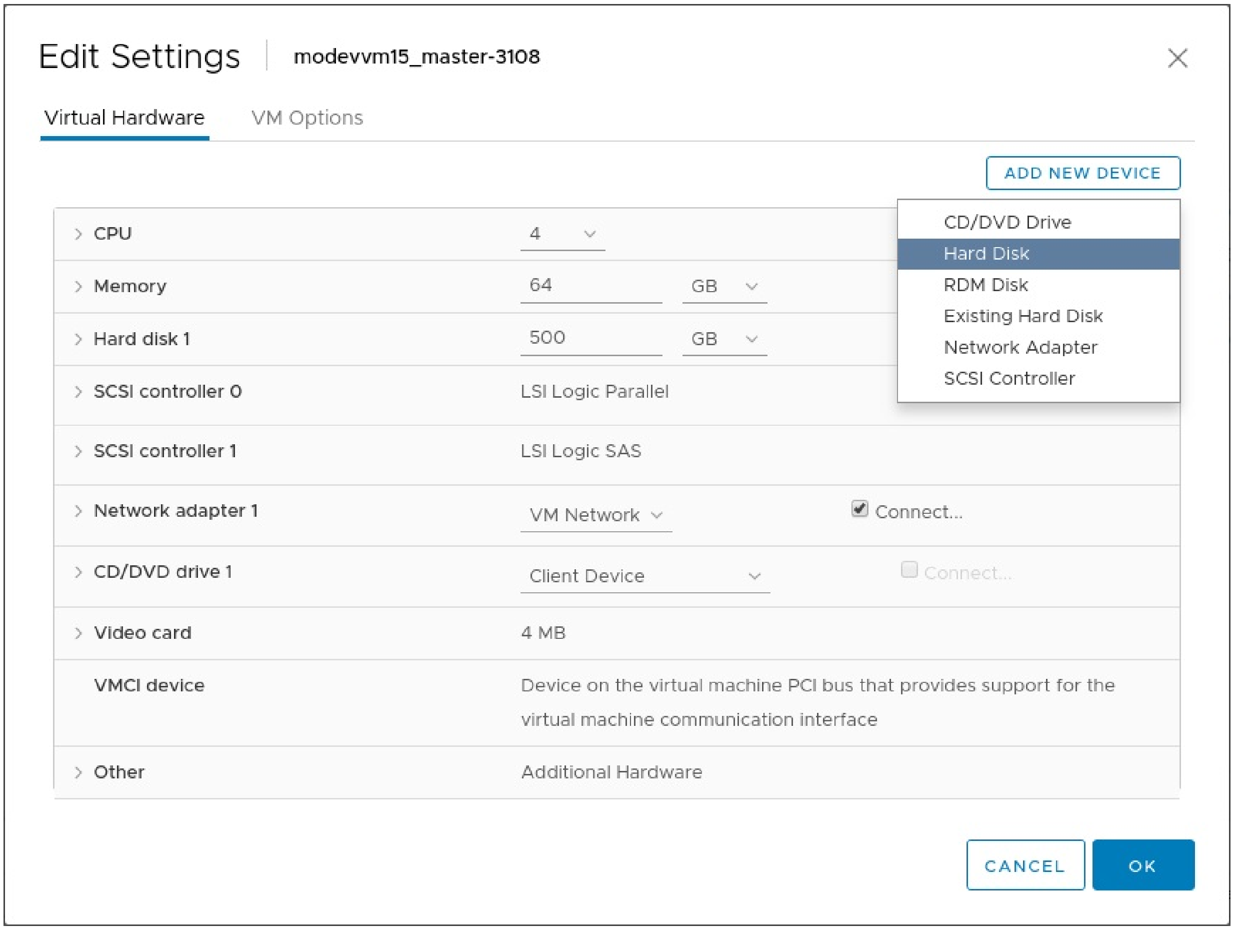
For a trial IBM Spectrum Discover deployment, a minimum of 100 GB capacity is required for the VMDK and 1 GB per 2 million indexed files can be used as a sizing metric if IBM Spectrum Discover agents are not to be configured. If IBM Spectrum Discover action agents are to be configured, a minimum of 100 GB capacity is required for the VMDK and 2 GB per 2 million index files can be used as a sizing metric.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**.



2. From the **ADD NEW DEVICE** drop-down menu in the upper-right hand corner of the dialog box, select **Hard Disk**.

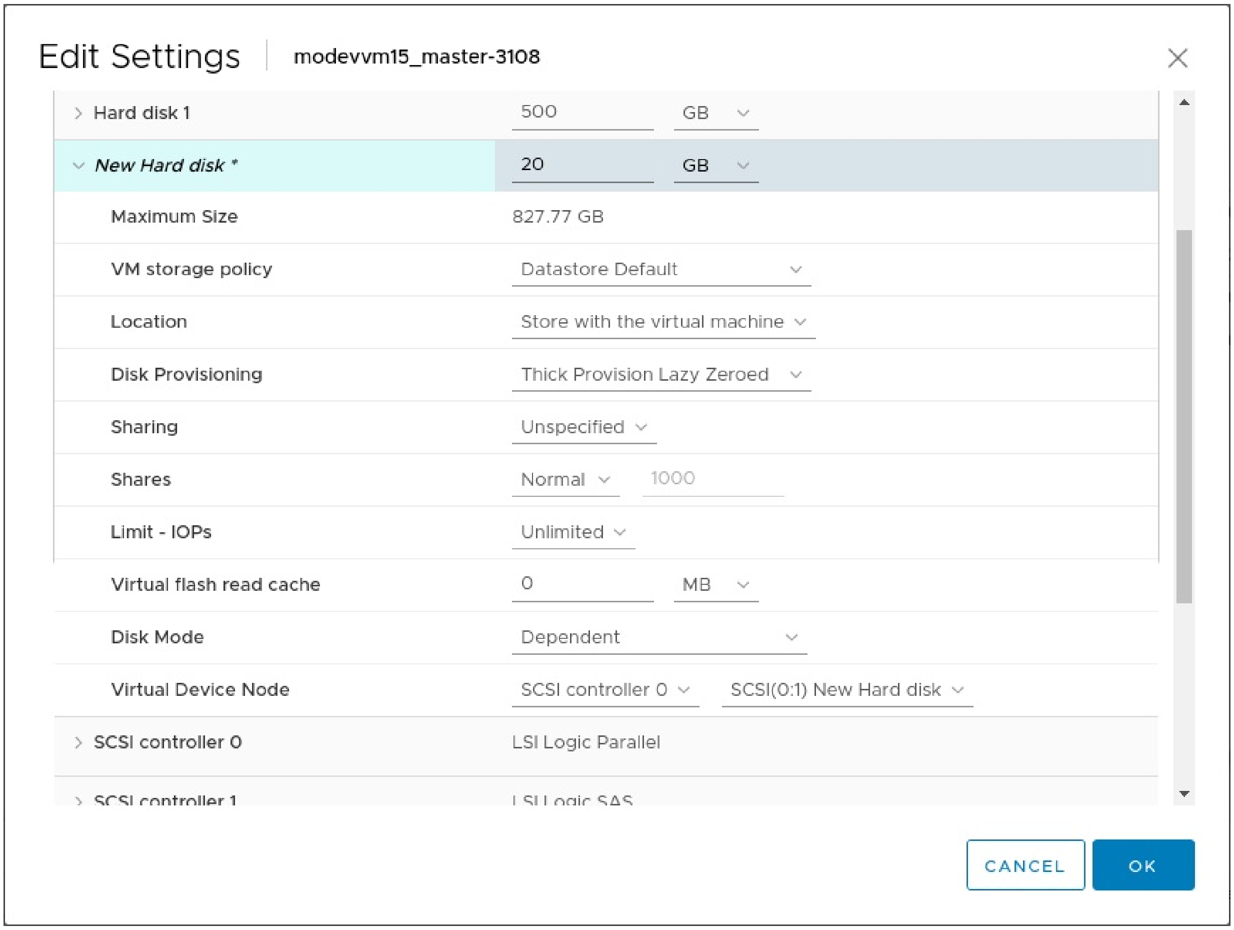


Edit Settings | modevvm15_master-3108

CPU	4		
Memory	64	GB	
Hard disk 1	500	GB	
SCSI controller 0	LSI Logic Parallel		
SCSI controller 1	LSI Logic SAS		
Network adapter 1	VM Network		☑ Connect...
CD/DVD drive 1	Client Device		□ Connect...
Video card	4 MB		
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface		
Other	Additional Hardware		

A **New Hard Disk** entry appears under **Virtual Hardware**.

3. Click **New Hard Disk** to expand the menu and select options for the disk.



Edit Settings | modevvm15_master-3108

Hard disk 1	500 GB	
New Hard disk *	20 GB	
Maximum Size	827.77 GB	
VM storage policy	Datastore Default	
Location	Store with the virtual machine	
Disk Provisioning	Thick Provision Lazy Zeroed	
Sharing	Unspecified	
Shares	Normal	1000
Limit - IOPs	Unlimited	
Virtual flash read cache	0 MB	
Disk Mode	Dependent	
Virtual Device Node	SCSI controller 0	SCSI(0:1) New Hard disk
SCSI controller 0	LSI Logic Parallel	
SCSI controller 1	LSI Logic SAS	

At this point, you can set the size, provisioning and location of the virtual disk. The default location is the datastore where the virtual appliance resides. If needed, you can select a different datastore

Note: Hard disk options shows an example of a new hard size of 20 GB. In practice this number should be much larger. For production environments, it is required to allocate more space for the persistent message queue. See the section for Planning in the IBM Spectrum Discover Knowledge Center.

4. Click **OK** to confirm your settings and create the virtual disk.

Adding a virtual disk for the IBM Spectrum Discover database

You can use the VMware vSphere Client to add the virtual disk required for IBM Spectrum Discover database to the virtual appliance.

Before you begin

Important:

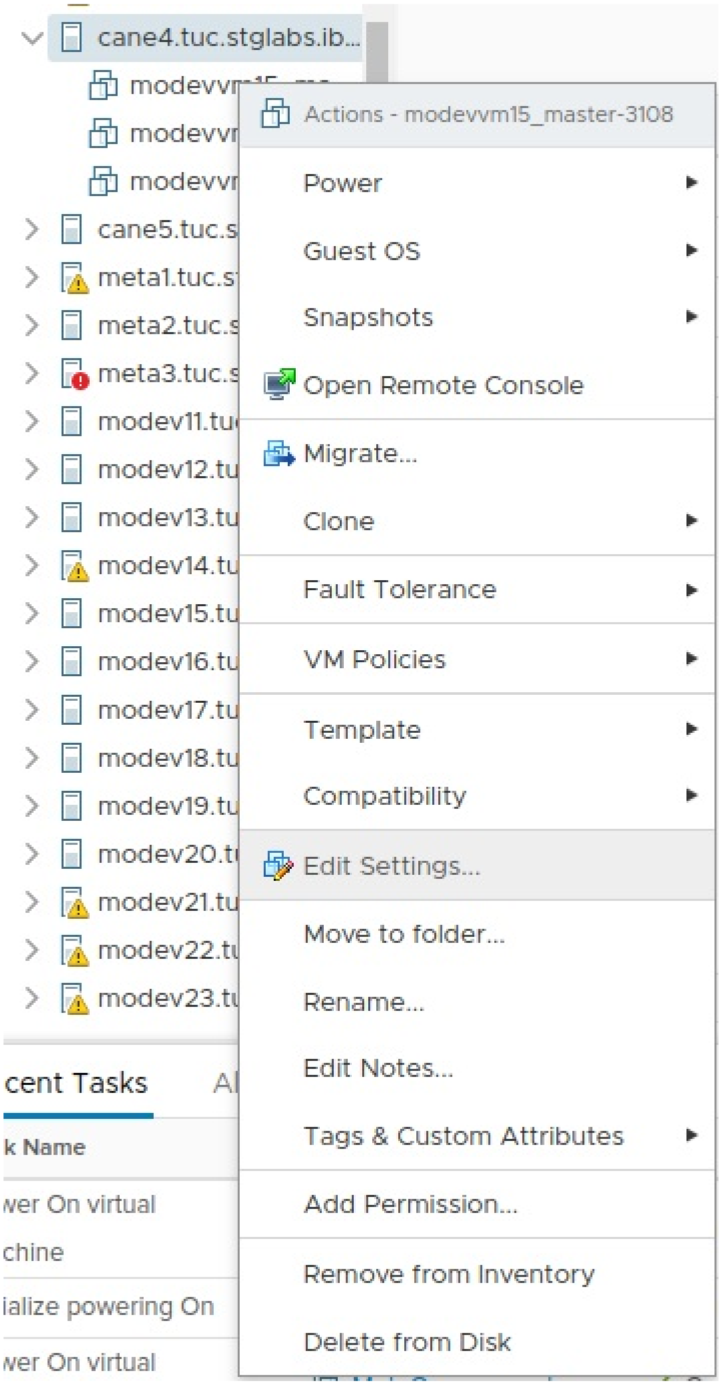
See the Planning section in the IBM Spectrum Discover Knowledge Center for detailed requirements for the database VMDK. For a single node production IBM Spectrum Discover deployment, a 2.5 TB thick provisioned, and lazy zeroed VMDK is required.

For a trial IBM Spectrum Discover deployment, a minimum of 100 GB capacity is required for the VMDK and 1 GB per 2 million indexed files can be used as a sizing metric if backup and restore is not required. If backup and restore is required, a minimum of 100 GB capacity is required for the VMDK and 2 GB per 2 million indexed files can be used as a sizing metric.

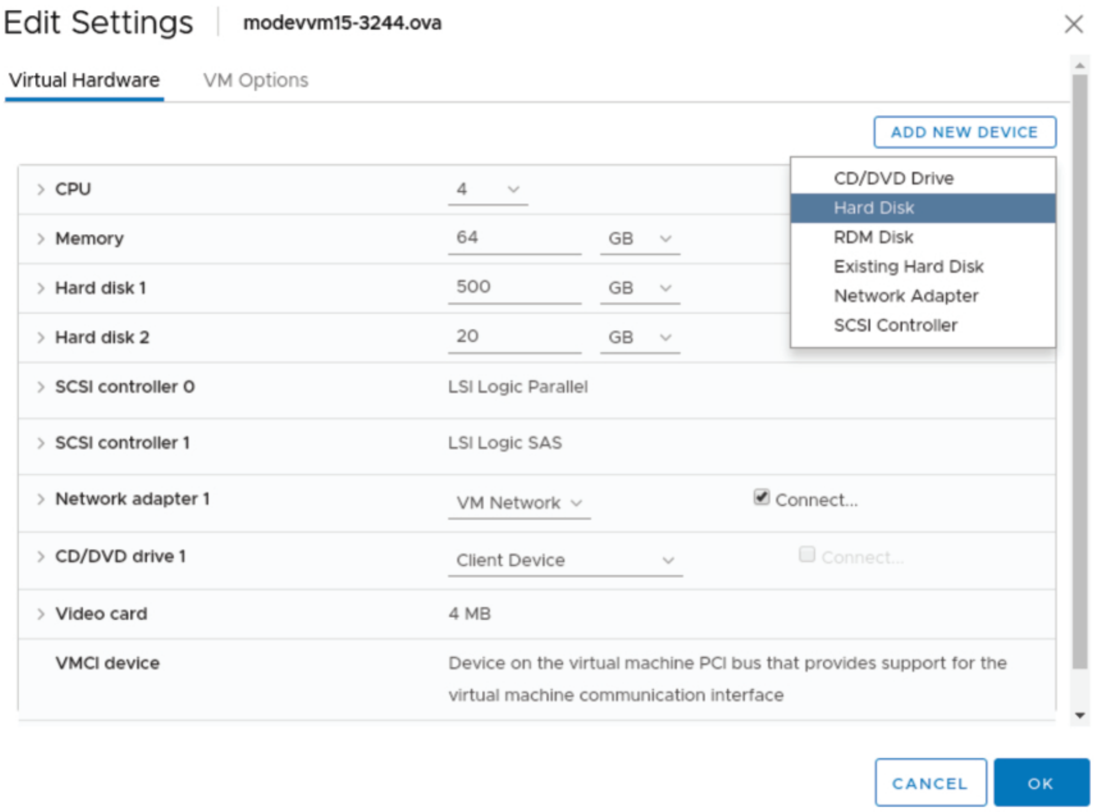
You can use the VMware vSphere Client to add the virtual disk required for the IBM Spectrum Discover database to the virtual appliance.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and then click **Edit Settings**.



2. From the **ADD NEW DEVICE** list, select **Hard Disk**.



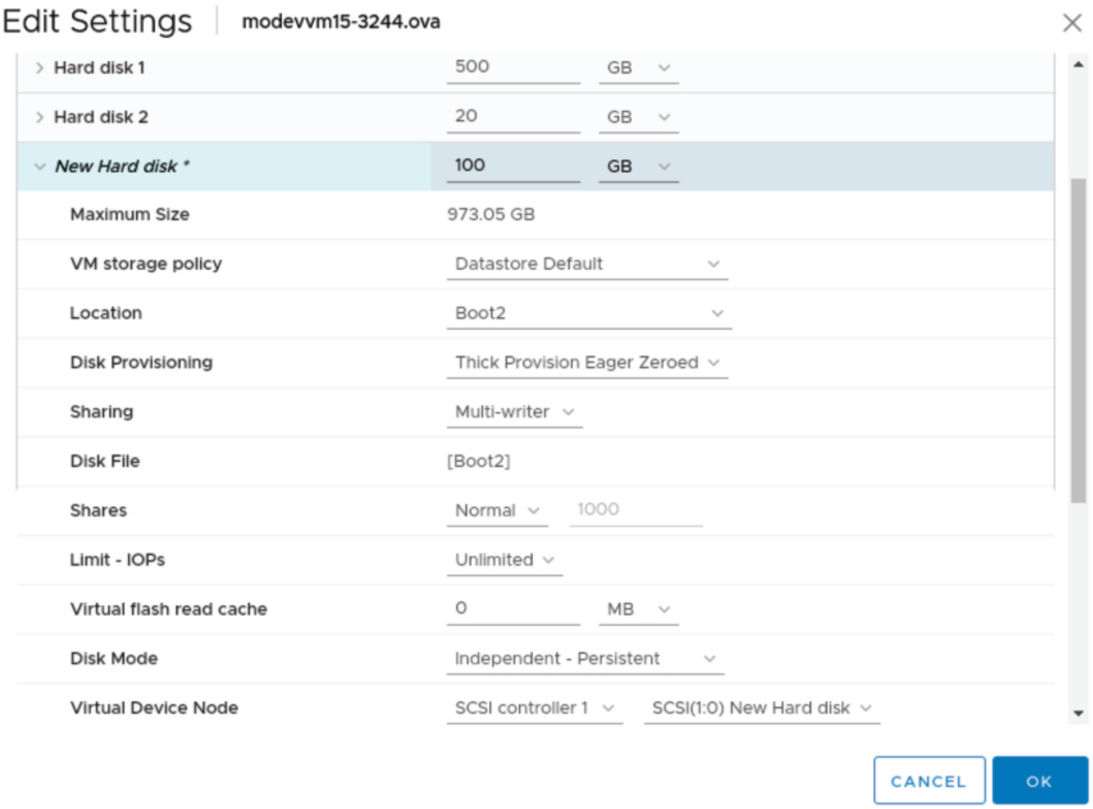
Edit Settings | modevvm15-3244.ova

CPU	4	
Memory	64 GB	
Hard disk 1	500 GB	
Hard disk 2	20 GB	
SCSI controller 0	LSI Logic Parallel	
SCSI controller 1	LSI Logic SAS	
Network adapter 1	VM Network	☑ Connect...
CD/DVD drive 1	Client Device	□ Connect...
Video card	4 MB	
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface	

pov00024

A **New Hard Disk** entry appears under **Virtual Hardware**.

3. Click **New Hard Disk** to expand the menu and select options for the disk.



Edit Settings | modevvm15-3244.ova

Hard disk 1	500	GB
Hard disk 2	20	GB
*New Hard disk **	100	GB
Maximum Size	973.05 GB	
VM storage policy	Datastore Default	
Location	Boot2	
Disk Provisioning	Thick Provision Eager Zeroed	
Sharing	Multi-writer	
Disk File	[Boot2]	
Shares	Normal	1000
Limit - IOPs	Unlimited	
Virtual flash read cache	0	MB
Disk Mode	Independent - Persistent	
Virtual Device Node	SCSI controller 1	SCSI(1:0) New Hard disk

pov00025

At this point, you can set the size, provisioning and location of the virtual disk. The default location is the data store where the virtual appliance resides. If needed, you can select a different datastore.

4. Click **OK** to confirm your settings and create the virtual disk.

Configuring CPU and memory allocation for the single node IBM Spectrum Discover virtual appliance

This section lists the step to increase the default allocations of CPU and memory for each IBM Spectrum Discover virtual appliance.

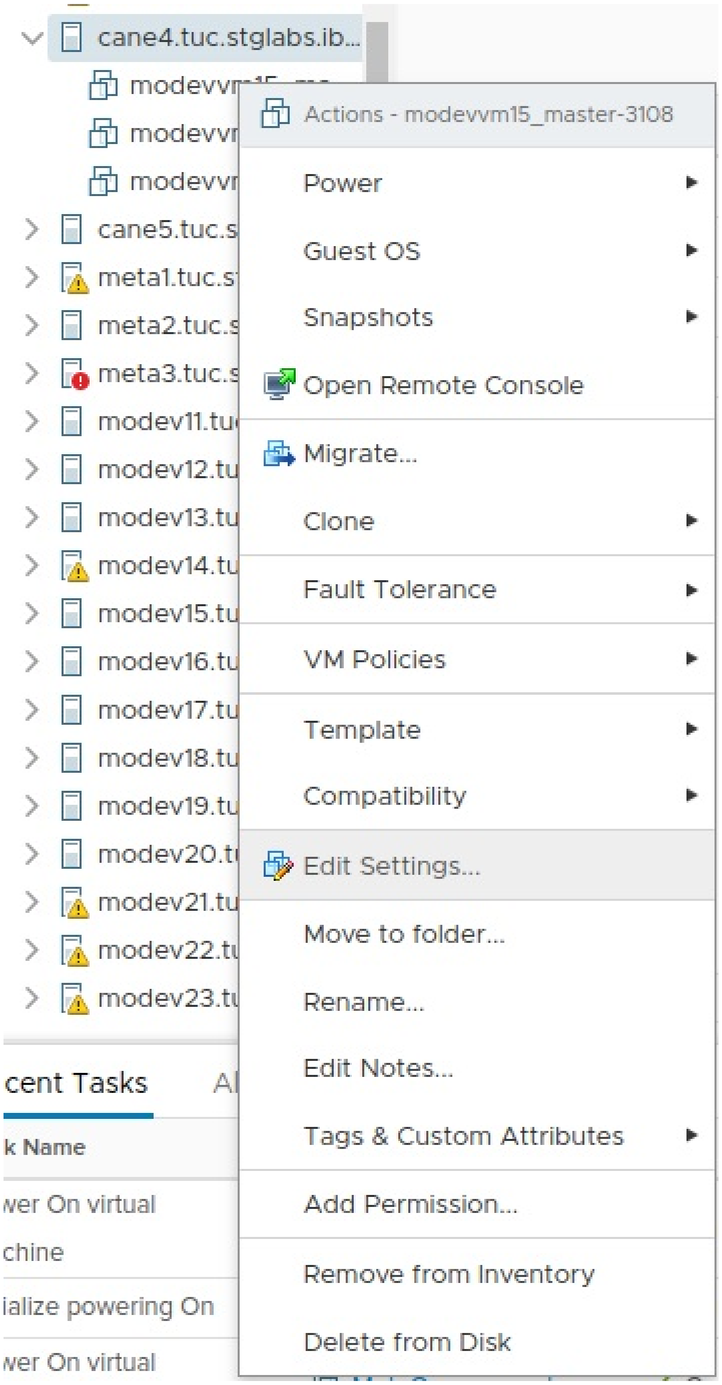
About this task

It is recommended to reserve the assigned memory assigned to the IBM Spectrum Discover virtual appliance to avoid running out of physical memory and swapping, which severely impacts database performance and stability.

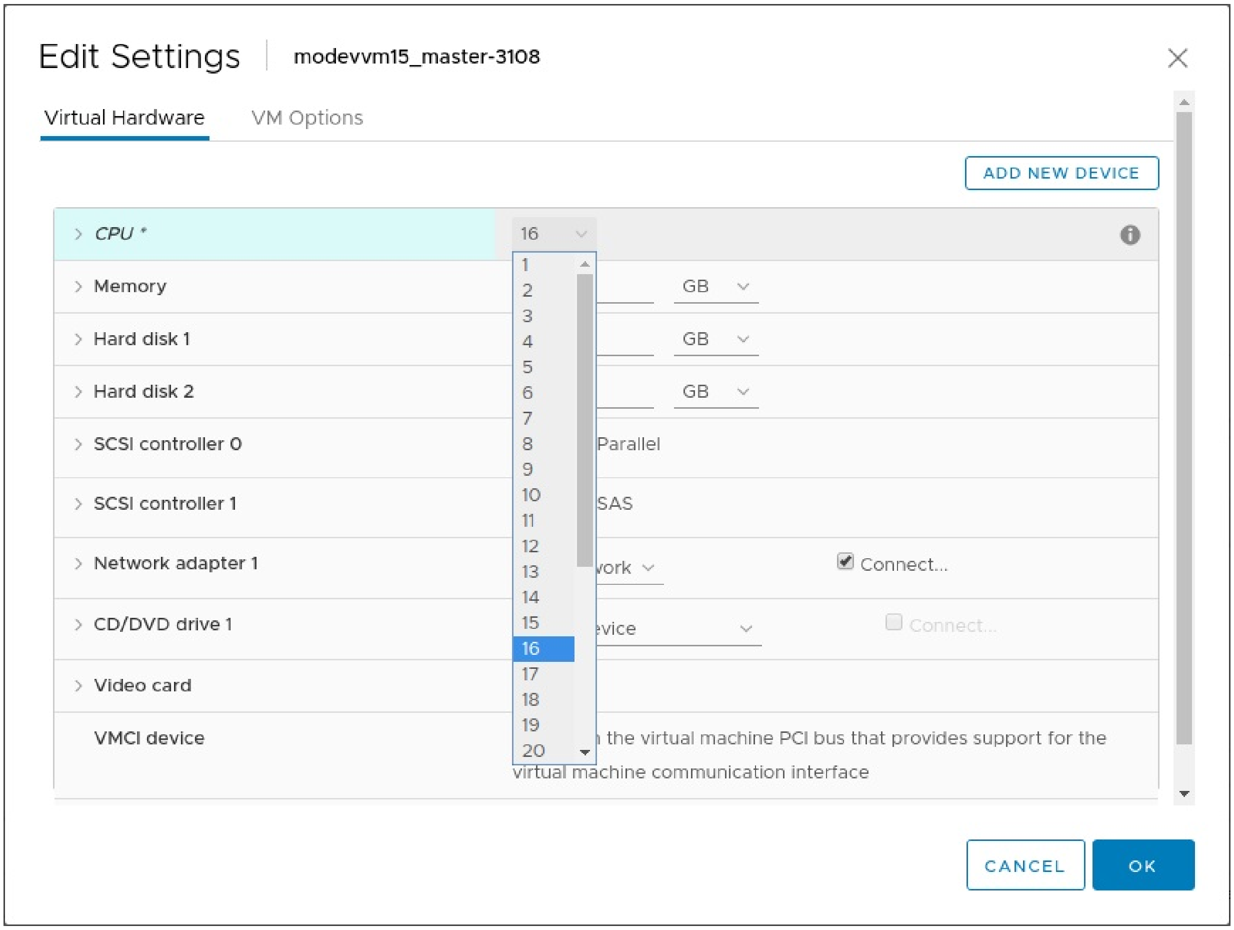
Important: A single node production IBM Spectrum Discover virtual appliance requires 128 GB RAM and 24 logical processors. 128 GB RAM and 24 logical processors is also recommended for the single node trial IBM Spectrum Discover virtual appliance. However, 64 GB and 8 local processors can be configured that support indexing of up to 25 million files and objects. See Planning for more details.

Procedure

1. In the vSphere client, right-click the IBM Spectrum Discover virtual appliance for which you want to change the CPU and memory allocation and click **Edit Settings**.



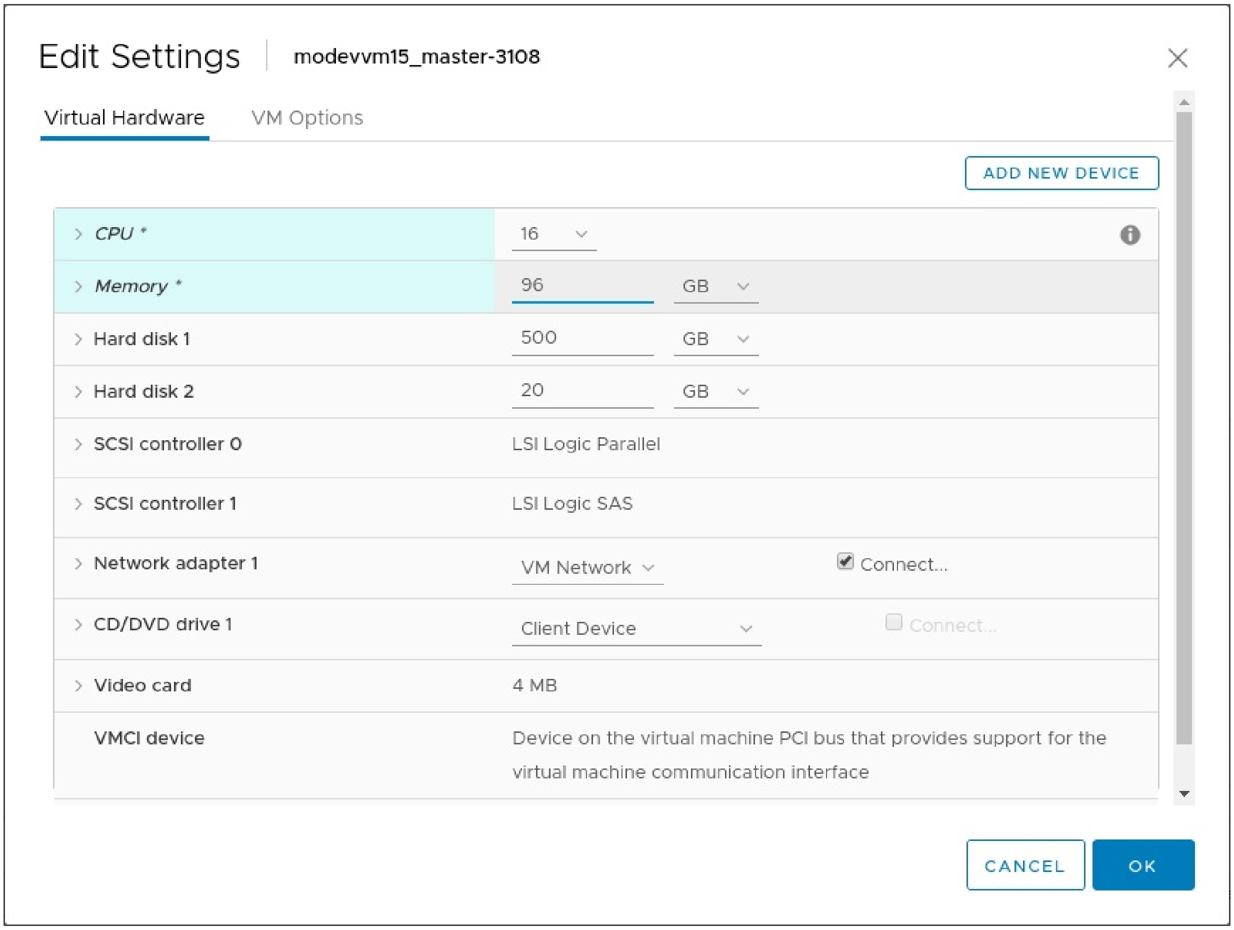
2. Under **Virtual Hardware**, from the **CPU** list, select the number that you want to increase the CPU allocation to.



CPU * 16

VMCI device: n the virtual machine PCI bus that provides support for the virtual machine communication interface

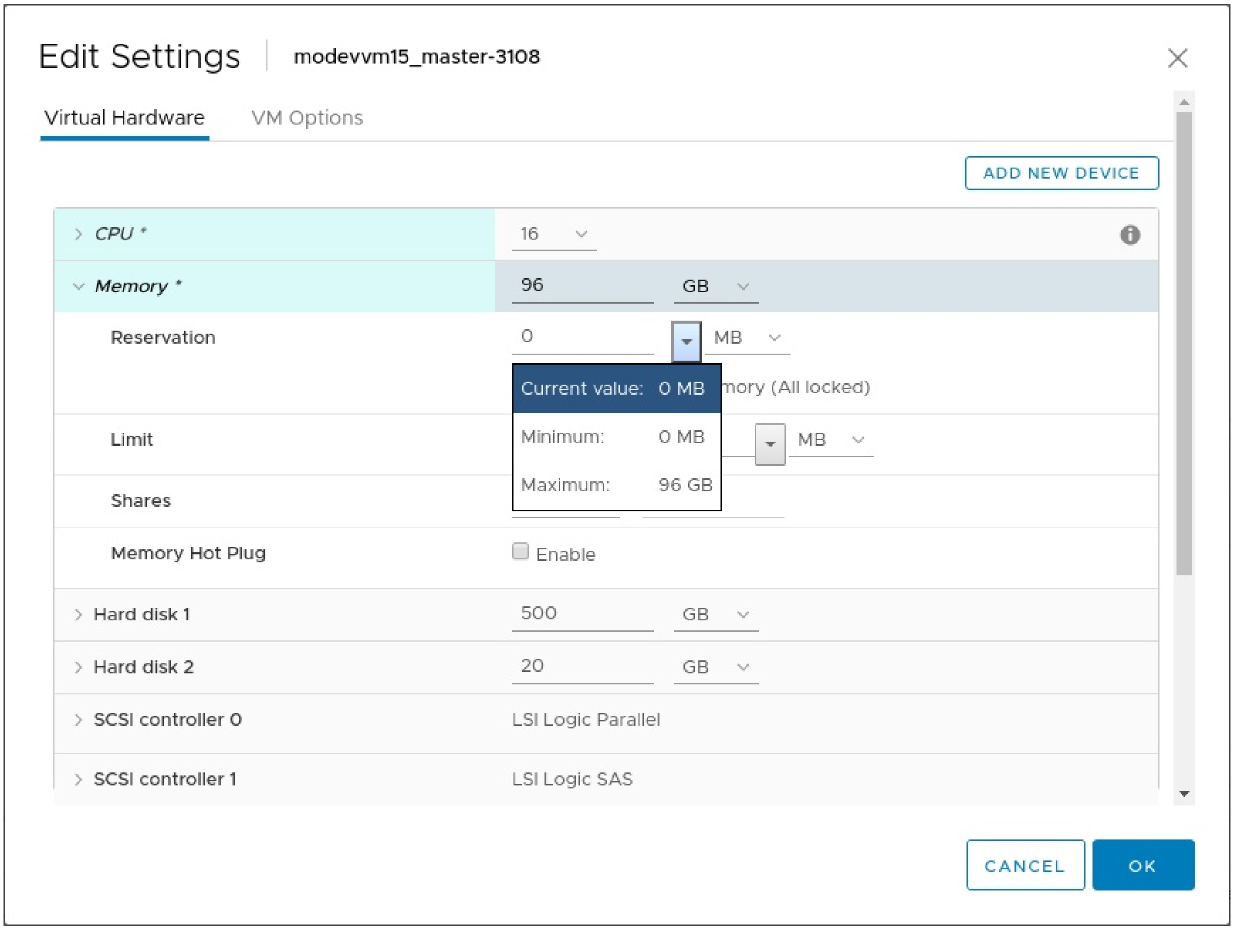
3. In the **Memory** field, enter the number that you want to change the memory allocation to and select the memory unit from the drop-down list.



Edit Settings | modevvm15_master-3108

CPU *	16	
Memory *	96 GB	
Hard disk 1	500 GB	
Hard disk 2	20 GB	
SCSI controller 0	LSI Logic Parallel	
SCSI controller 1	LSI Logic SAS	
Network adapter 1	VM Network	☑ Connect...
CD/DVD drive 1	Client Device	□ Connect...
Video card	4 MB	
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface	

4. In the **Reservation** field under **Memory**, change the number according to the changed memory allocation and select the memory unit from the drop-down list.



modevvm15_master-3108

CPU * 16

Memory * 96 GB

Reservation 0 MB

Current value: 0 MB

Minimum: 0 MB

Maximum: 96 GB

mory (All locked)

Limit MB

Shares

Memory Hot Plug □ Enable

Hard disk 1 500 GB

Hard disk 2 20 GB

SCSI controller 0 LSI Logic Parallel

SCSI controller 1 LSI Logic SAS

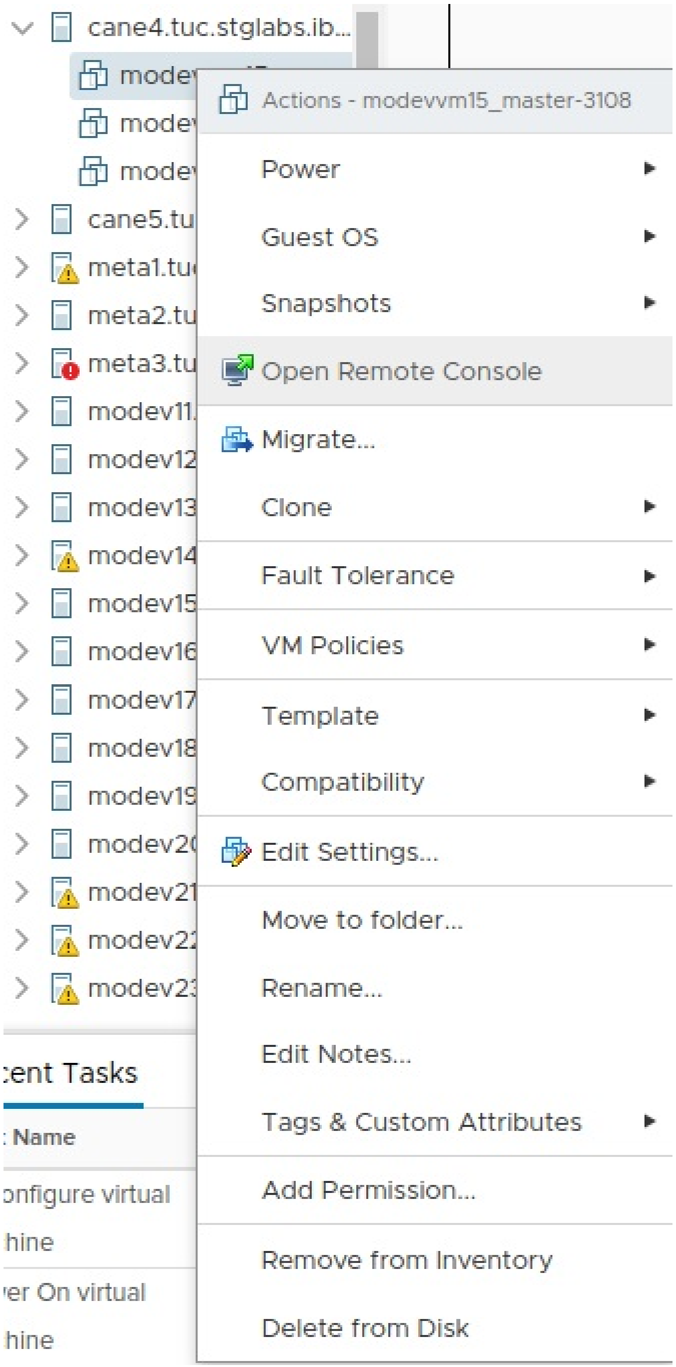
5. Click **OK** to confirm the changes in CPU and memory allocation.

Configuring networking and perform provisioning of a single node trial or single node production IBM Spectrum Discover virtual appliance

After virtual appliance in the IBM Spectrum Discover is deployed, and storage, CPU, and memory are configured, you need to configure networking and then provision the virtual appliances by using a provisioning tool.

Procedure

1. Power on the virtual appliance.
2. In the vSphere Client, right-click the virtual appliance and click **Open Remote Console**.



3. At the virtual appliance login prompt, enter the user name and the password to log in. The default user name is **moadmin** and the default password is **Passw0rd**.
4. Change the directory to `/opt/ibm/metaocean/configuration`.

```
cd /opt/ibm/metaocean/configuration
```

Note: IBM Spectrum Discover requires a fully qualified domain name (FQDN) that is registered in a customer supplied domain name server (DNS). The customer supplied FQDN must be resolvable by the customer supplied DNS from the IBM Spectrum Discover node for the IBM Spectrum Discover virtual appliance to operate properly

5. Configure the IBM Spectrum Discover virtual appliance networking settings by running the following command:

```
sudo ./mmconfigappliance
```

Note: One or more nodes IBM Spectrum Discover must be able to communicate with a customer supplied network time protocol (NTP) server in order to operate properly.

The following table lists the definitions of the required settings:

Parameter	Value format	Recommended value	Example
HostName	host.domain.com	The fully qualified domain name of the node	node1234.example.com
Interface	ensXXX	The Ethernet interface to use for the virtual appliance networking.	ens192
IPAddress	xxx.xxx.xxx.xxx	The IP address of the node.	10.10.200.10
NetMask	xxx.xxx.xxx.xxx	The network mask for the IP range of the node.	255.255.254.0
Gateway	host.domain.com	The IP address of the network gateway.	10.10.200.1
DNS	ensXXX	The IP address of a single DNS server.	10.10.200.35
NTPServer	xxx.xxx.xxx.xxx or host.domain.com	The fully qualified domain or the IP address of the NTP server.	pool1.ntp.org
Mode	single or multi	single	single

Note: [During this step, you are prompted to:

- Set the timezone. To change the timezone to something other than UTC, go through the prompts to choose your continent, country, or region. (No other setup is required to set the timezone.)
- Change the `moadmin` password.

]

[The **`mmconfigappliance`** process takes approximately 1 hour to complete for a production system.]

[Check for the Kubernetes `network_cidr` and `service_cluster_ip_range` values. These specified values must not conflict with the existing host network IP data. The default values are:

```
network_cidr: 10.1.0.0/16
service_cluster_ip_range: 10.0.0.0/16
```

]

[If you are prompted for the `network_cidr` or `service_cluster_ip_range`, consider this information to avoid network conflicts. Private network ranges that might be a good choice are:

- `10.0.0.0` to `10.255.255.255`
- `172.16.0.0` to `172.31.255.255`

]

[For example, you can enter:

- `172.31.0.0/16` for the `network_cidr`

- `172.30.0.0/16` for the `service_cluster_ip_range`

```
network_cidr: 172.31.0.0/16
service_cluster_ip_range: 172.30.0.0/16
```

IBM Spectrum Discover checks to see whether the system host IP overlaps with the Kubernetes default network and service IP range. If this overlap is detected, you get an error message similar to this:

```
Host network (10.1.10.10) is overlapped with the default Kubernetes network (10.1.0.0/16).
Please enter in a new value for the Kubernetes network.
```

Upon successful completion, an output similar to the following is displayed.

```
PLAY RECAP *********************************************************************
**********************************************************************
**************************************
203.0.113.14        : ok=241  changed=209  unreachable=0    failed=0
canevm7.example.com : ok=9    changed=6    unreachable=0    failed=0
canevm8.example.com : ok=9    changed=6    unreachable=0    failed=0
canevm9.example.com : ok=9    changed=6    unreachable=0    failed=0
```

The process is completed successfully when you do not see any messages that say `failed` or when you see a message that the failed count = 0.

For more information on using Kubernetes settings, see: https://www.ibm.com/support/knowledgecenter/en/SSBS6K_3.1.2/installing/config_yaml.html

Known issues with deploying and configuring for single node

In case of any errors that you might encounter while deploying or configuring IBM Spectrum Discover, review the following information for details and possible workarounds.

Issue	Description	Resolution or workaround

Log shows error after deployment	Even after successful deployment, log might show some errors with messages similar to the following:
	`2018-10-15 11:33:00,557 p=12895 u=root |`
	`fatal: [203.0.113.18]:`
	`FAILED! => {"changed": false, "cmd": "awk '/`
	`sse4_2/ {exit 42}'`
	`/proc/cpuinfo", "delta": "0:00:00.004105",`
	`"end":`
	`"2018-10-15 11:33:00.540782", "failed":`
	`true, "rc": 42,`
	`"start": "2018-10-15 11:33:00.536677",`
	`"stderr": "",`
	`"stderr_lines": [], "stdout": "",`
	`"stdout_lines": []}`
	`2018-10-15 11:33:00,558 p=12895 u=root`
	`| ...ignoring`
		These error messages can be ignored.

Deploy and configure a multi-node production IBM Spectrum Discover appliance cluster

This section provides information on how to deploy and configure IBM Spectrum Discover multi-node production virtual appliance.

Deploying a multi-node production IBM Spectrum Discover virtual appliance cluster

The IBM Spectrum Discover software is available as an OVA (open virtualization appliance) file. You can deploy it on your VMware ESXi server by using the VMware vSphere Client:

Before you begin

- Download the IBM Spectrum Discover OVA file on the local system or obtain the URL to an IBM Spectrum Discover OVA file accessible on the internet.
- Review the deployment and configuration known issues and workarounds. For more information, see "Known issues with deploying and configuring for single node" on page 45.
- For a multi-node production IBM Spectrum Discover cluster requires three virtual nodes - one master node and two worker nodes.

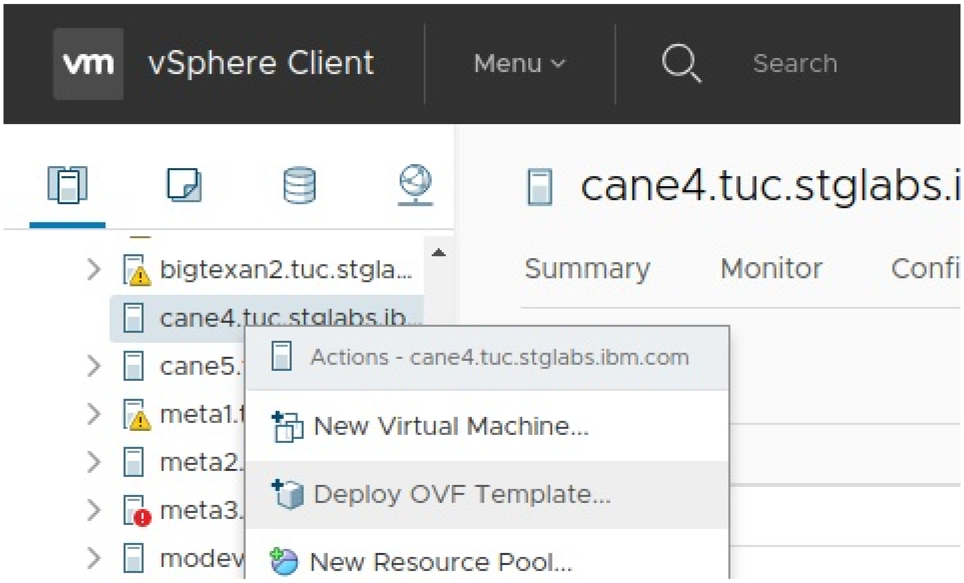
Important: You cannot use the HTML5 vSphere client to deploy the OVF image.

About this task

Deploy the IBM Spectrum Discover virtual appliance as follows by using the **Deploy OVF Template** wizard of the VMware vSphere Client.

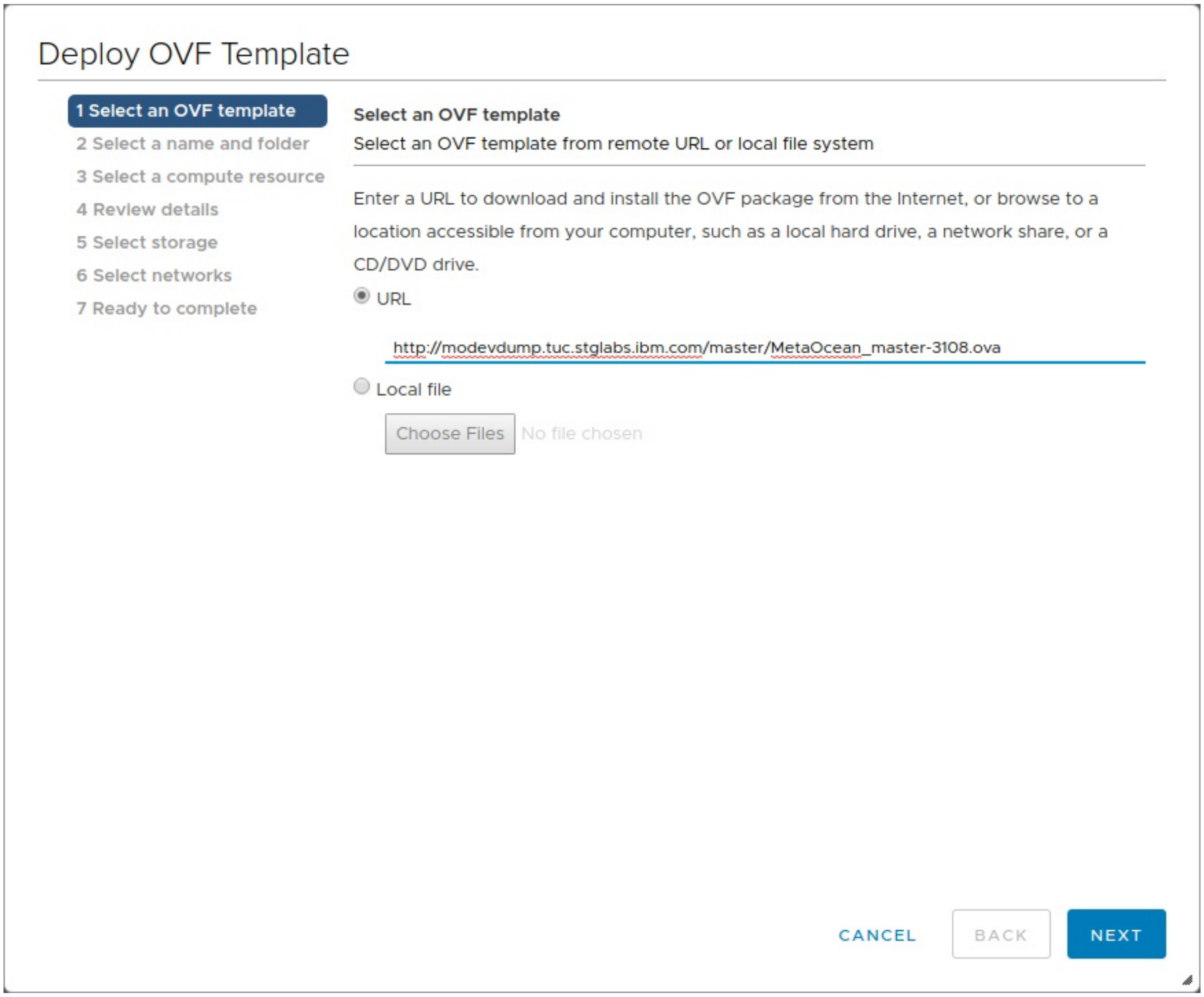
Procedure

1. In the vSphere Client, right-click the ESXi server on which you want to deploy the virtual appliance and click **Deploy OVF Template**.



   The **Deploy OVF Template** wizard appears.
2. Select the IBM Spectrum Discover virtual appliance that you want to deploy and click **Next**.

You can either select an OVA file that you have downloaded on the local system or you can specify a URL to the OVA file.



Deploy OVF Template

Select an OVF template

Select an OVF template from remote URL or local file system

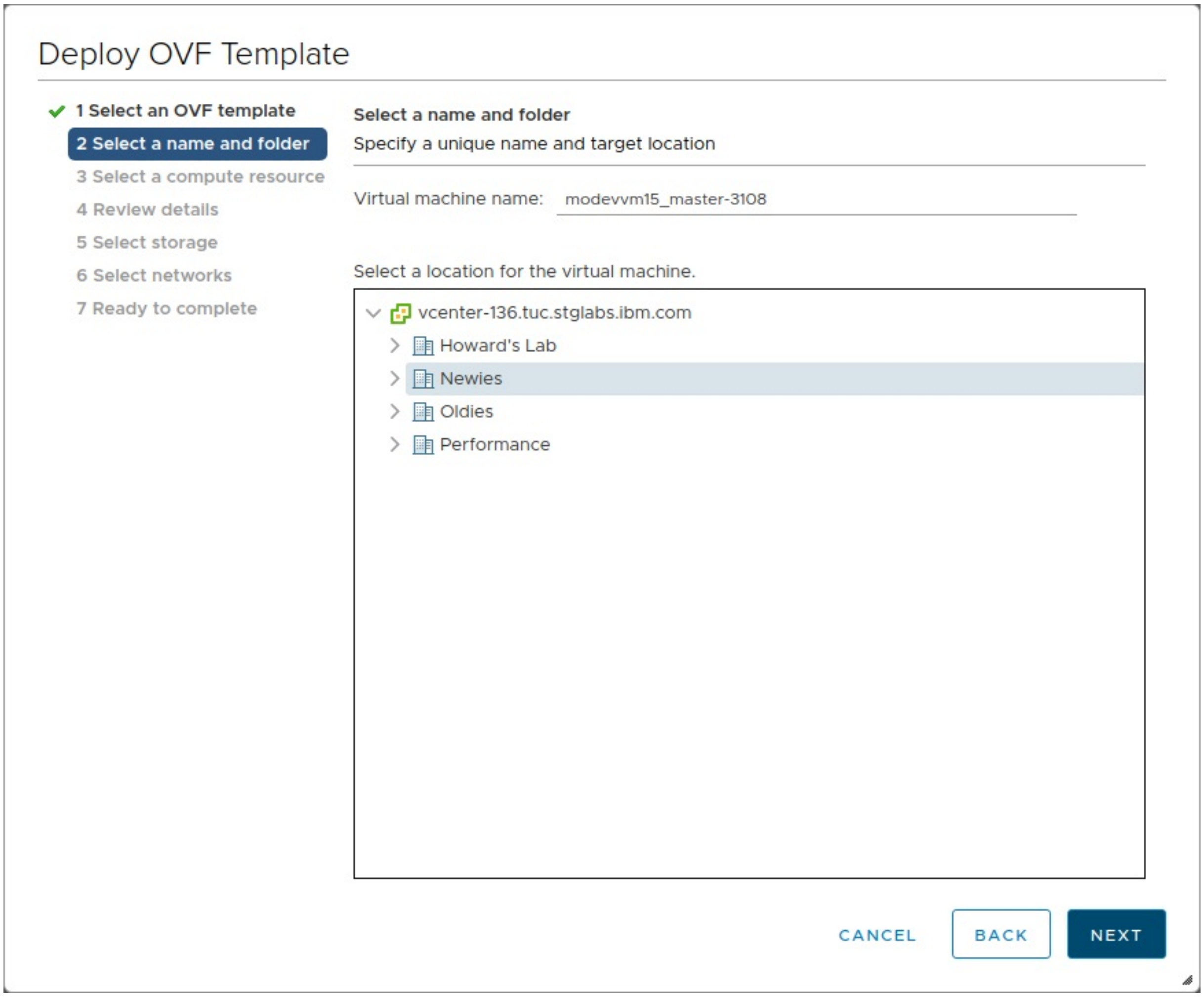
Enter a URL to download and install the OVF package from the Internet, or browse to a location accessible from your computer, such as a local hard drive, a network share, or a CD/DVD drive.

● URL

http://modevdump.tuc.stglabs.ibm.com/master/MetaOcean_master-3108.ova

○ Local file

3. Specify the name of the virtual appliance or accept the default name and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template

2 Select a name and folder

3 Select a compute resource

4 Review details

5 Select storage

6 Select networks

7 Ready to complete

Select a name and folder

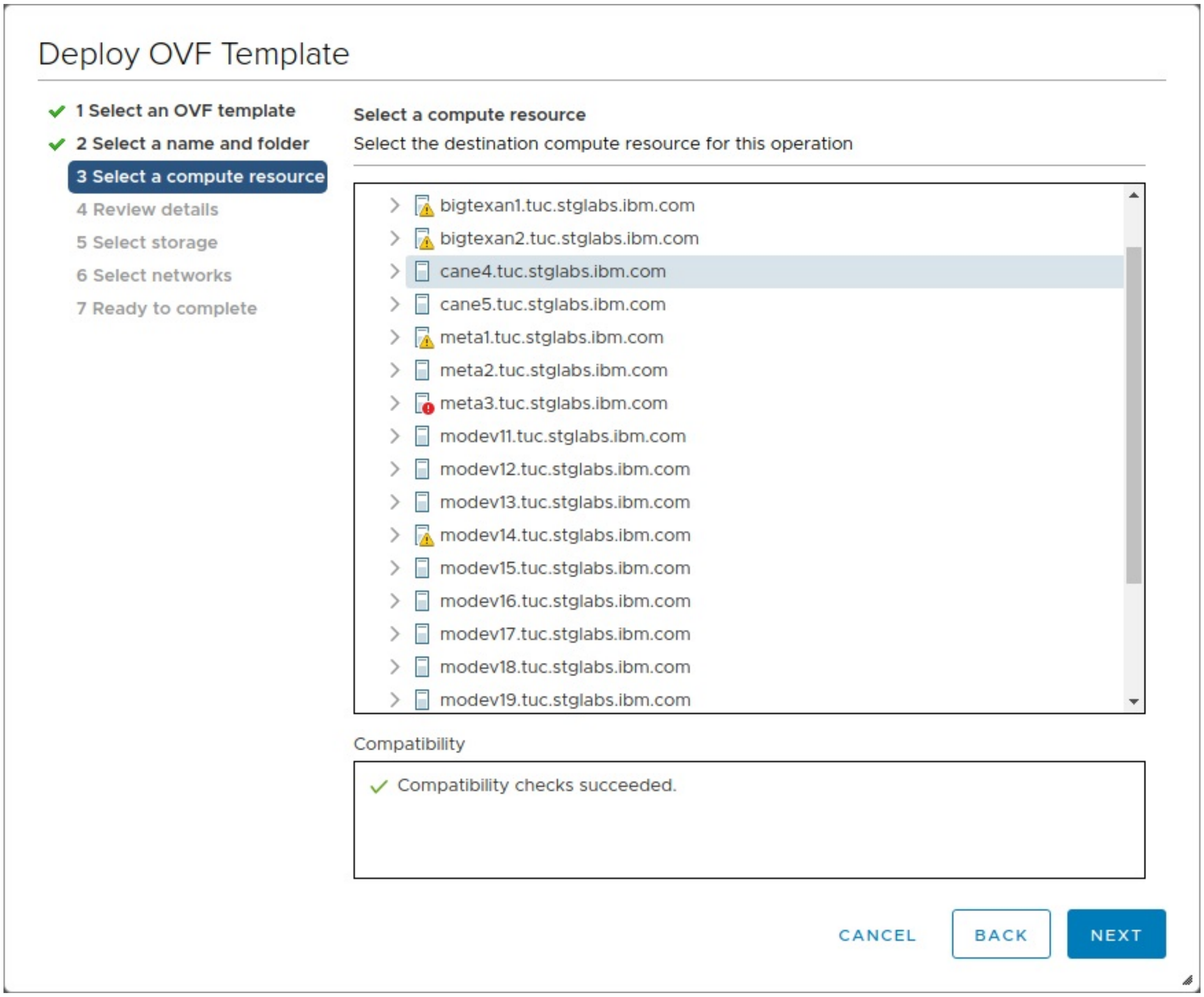
Specify a unique name and target location

Virtual machine name: modevvm15_master-3108

Select a location for the virtual machine.

- vcenter-136.tuc.stglabs.ibm.com
  - Howard's Lab
  - Newies
  - Oldies
  - Performance

4. Select the physical server on which you want to deploy the virtual appliance, and click **Next**.



Deploy OVF Template

- ✓ 1 Select an OVF template
- ✓ 2 Select a name and folder
- 3 Select a compute resource
- 4 Review details
- 5 Select storage
- 6 Select networks
- 7 Ready to complete

Select a compute resource

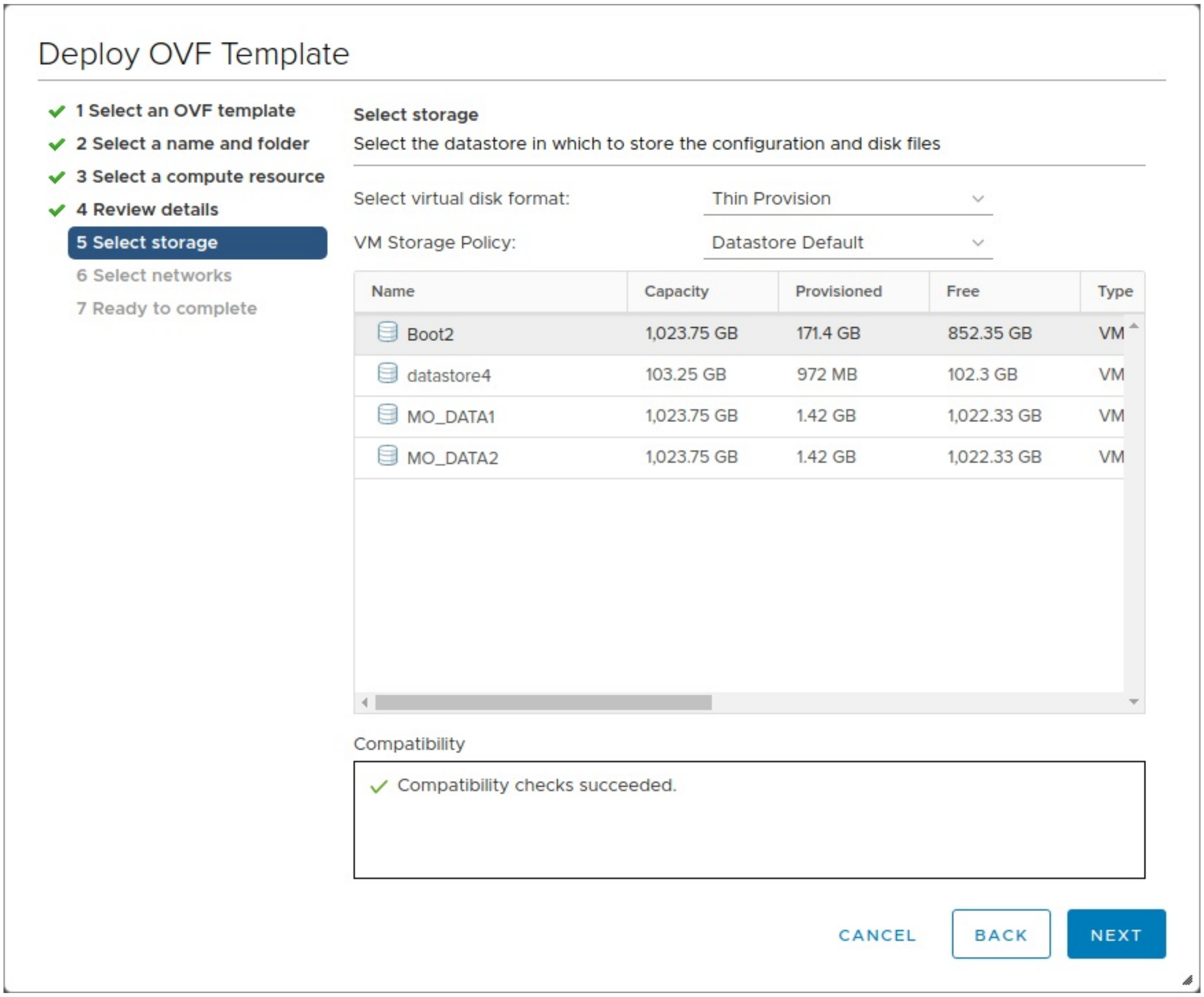
Select the destination compute resource for this operation

- bigtexan1.tuc.stglabs.ibm.com
- bigtexan2.tuc.stglabs.ibm.com
- cane4.tuc.stglabs.ibm.com
- cane5.tuc.stglabs.ibm.com
- meta1.tuc.stglabs.ibm.com
- meta2.tuc.stglabs.ibm.com
- meta3.tuc.stglabs.ibm.com
- modev11.tuc.stglabs.ibm.com
- modev12.tuc.stglabs.ibm.com
- modev13.tuc.stglabs.ibm.com
- modev14.tuc.stglabs.ibm.com
- modev15.tuc.stglabs.ibm.com
- modev16.tuc.stglabs.ibm.com
- modev17.tuc.stglabs.ibm.com
- modev18.tuc.stglabs.ibm.com
- modev19.tuc.stglabs.ibm.com

Compatibility

✓ Compatibility checks succeeded.

5. Review the details and click **Next**.
6. Select the check box to accept the terms of the licenses and click **Next**.
7. Select the data store for the virtual appliance and the virtual disk format, and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template
✓ 2 Select a name and folder
✓ 3 Select a compute resource
✓ 4 Review details
5 Select storage
6 Select networks
7 Ready to complete

Select storage

Select the datastore in which to store the configuration and disk files

Select virtual disk format: Thin Provision

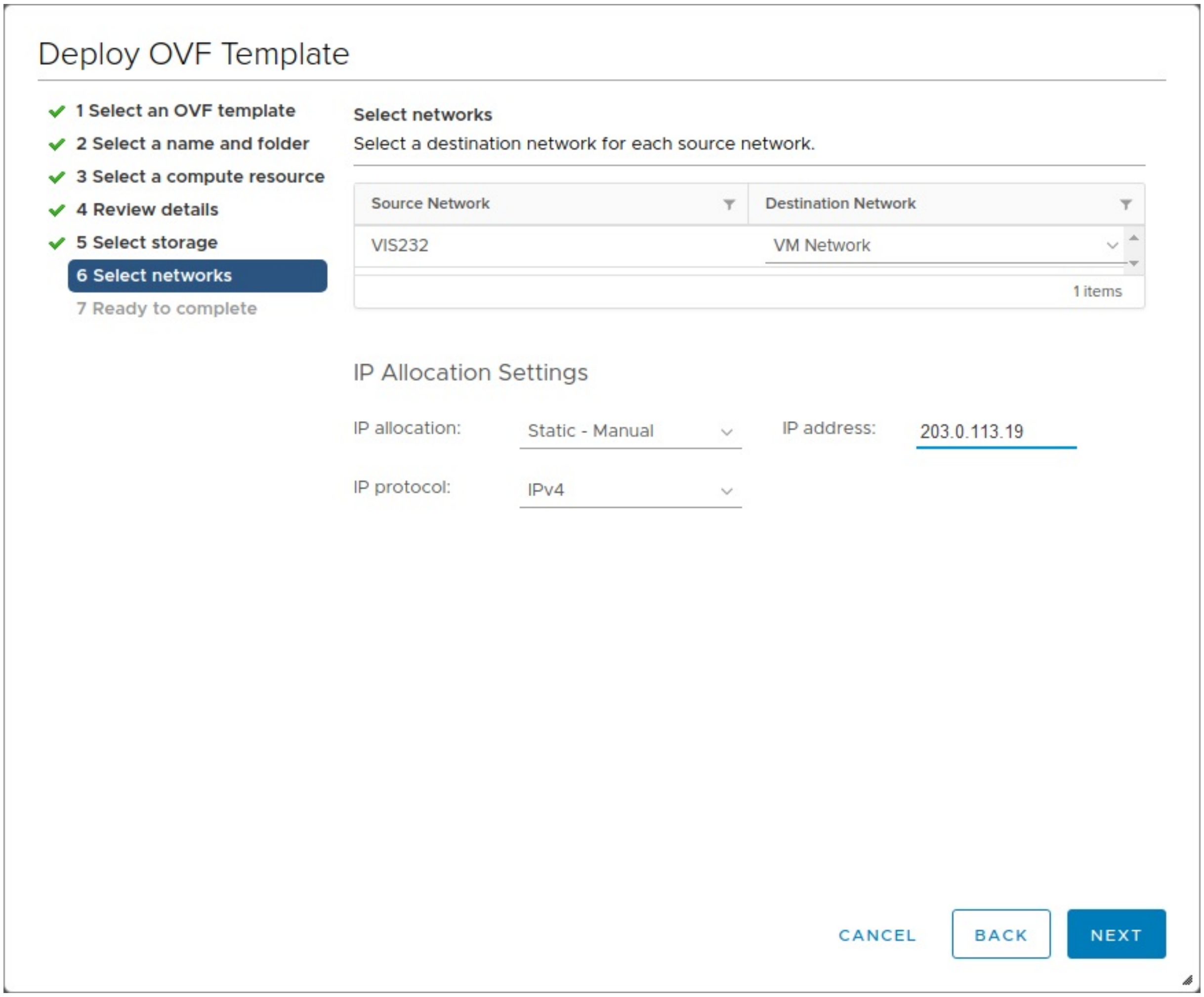
VM Storage Policy: Datastore Default

Name	Capacity	Provisioned	Free	Type
Boot2	1,023.75 GB	171.4 GB	852.35 GB	VM
datastore4	103.25 GB	972 MB	102.3 GB	VM
MO_DATA1	1,023.75 GB	1.42 GB	1,022.33 GB	VM
MO_DATA2	1,023.75 GB	1.42 GB	1,022.33 GB	VM

Compatibility

✓ Compatibility checks succeeded.

8. Select the VM network for the virtual appliance and click **Next**.



Deploy OVF Template

✓ 1 Select an OVF template
✓ 2 Select a name and folder
✓ 3 Select a compute resource
✓ 4 Review details
✓ 5 Select storage
6 Select networks
7 Ready to complete

Select networks

Select a destination network for each source network.

Source Network	Destination Network
VIS232	VM Network

1 items

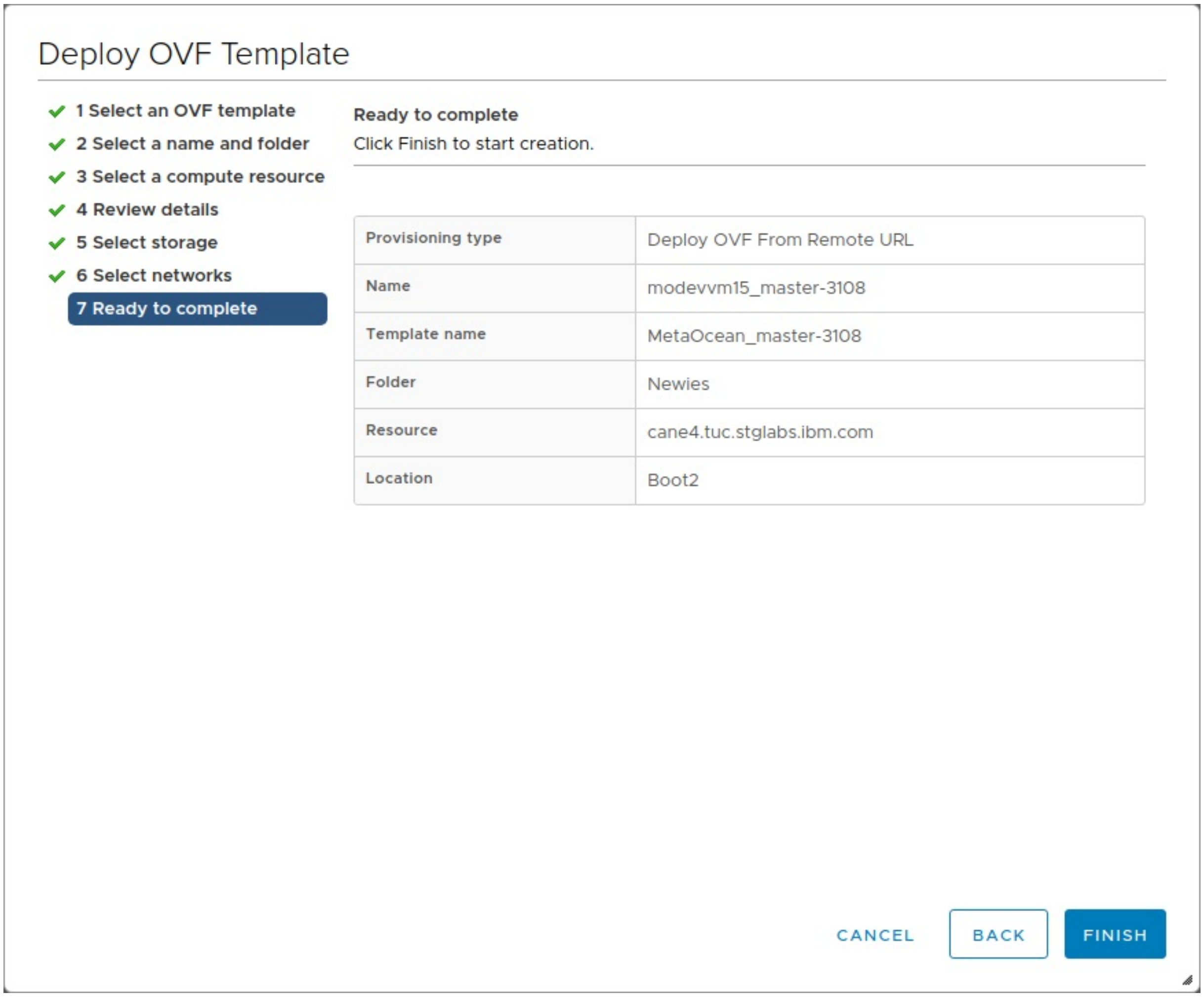
IP Allocation Settings

IP allocation: Static - Manual

IP address: 203.0.113.19

IP protocol: IPv4

9. Review the settings and click **Finish**.



Ready to complete

Click Finish to start creation.

Provisioning type	Deploy OVF From Remote URL
Name	modevvm15_master-3108
Template name	MetaOcean_master-3108
Folder	Newies
Resource	cane4.tuc.stglabs.ibm.com
Location	Boot2

The IBM Spectrum Discover virtual node gets created and the storage is provisioned.

Note: Do not power on the virtual appliance until storage, CPU, and memory have been configured.

10. Repeat step "1" on page 46 through step "9" on page 51 for each of the three nodes in the IBM Spectrum Discover multi-node production virtual appliance cluster.

Configuring storage for a multi-node production IBM Spectrum Discover virtual appliance cluster

Each node in the IBM Spectrum Discover three-node production virtual appliance cluster requires a VMDK storage device for the persistent message queue and also requires three shared VMDK storage device for the database.

Before you begin

Note: The persistent message queue and the three shared VMDK storage devices for the database are in addition to the base OS and software VMDK that was automatically configured during the initial IBM Spectrum Discover virtual appliance deployment.

Procedure

1. Do the following steps on the first IBM Spectrum Discover virtual appliance node in the three-node IBM Spectrum Discover cluster:

   a) Add a virtual disk for the IBM Spectrum Discover persistent message queues.

      For more information, see "Adding virtual disks for IBM Spectrum Discover persistent message queues for the first node in the multi-node cluster" on page 53.

b) Add an LSI Logic Storage Controller to manage the shared virtual disks that are required for the IBM Spectrum Discover database.

For more information see "Adding an LSI Logic SCSI Controller to the first virtual appliance node on the IBM Spectrum Discover cluster" on page 56

c) Add the three shared virtual disks for the IBM Spectrum Discover database to the first virtual appliance in the IBM Spectrum Discover cluster.

For more information, see "Adding virtual shared disks for the database to the first node in the IBM Spectrum Discover cluster." on page 59.

2. Do the following steps on the second and third IBM Spectrum Discover virtual appliances in the IBM Spectrum Discover cluster:

a) Add a virtual disk for the IBM Spectrum Discover persistent message queues.

For more information, see "Adding a virtual disk for IBM Spectrum Discover persistent message queues for the second and third nodes in the multi-node cluster" on page 62

b) Add an LSI Logic Storage Controller to manage the shared virtual disks that are required for the IBM Spectrum Discover database.

For more information, see "Adding LSI Logic SCSI Controller to the second and third virtual appliance nodes in the IBM Spectrum Discover cluster" on page 64

c) Add the three shared virtual disks for the IBM Spectrum Discover database to the remaining two nodes in the IBM Spectrum Discover cluster.

For more information, see "Adding shared virtual disks for the second and third nodes in the IBM Spectrum Discover cluster" on page 68

Adding virtual disks for IBM Spectrum Discover persistent message queues for the first node in the multi-node cluster

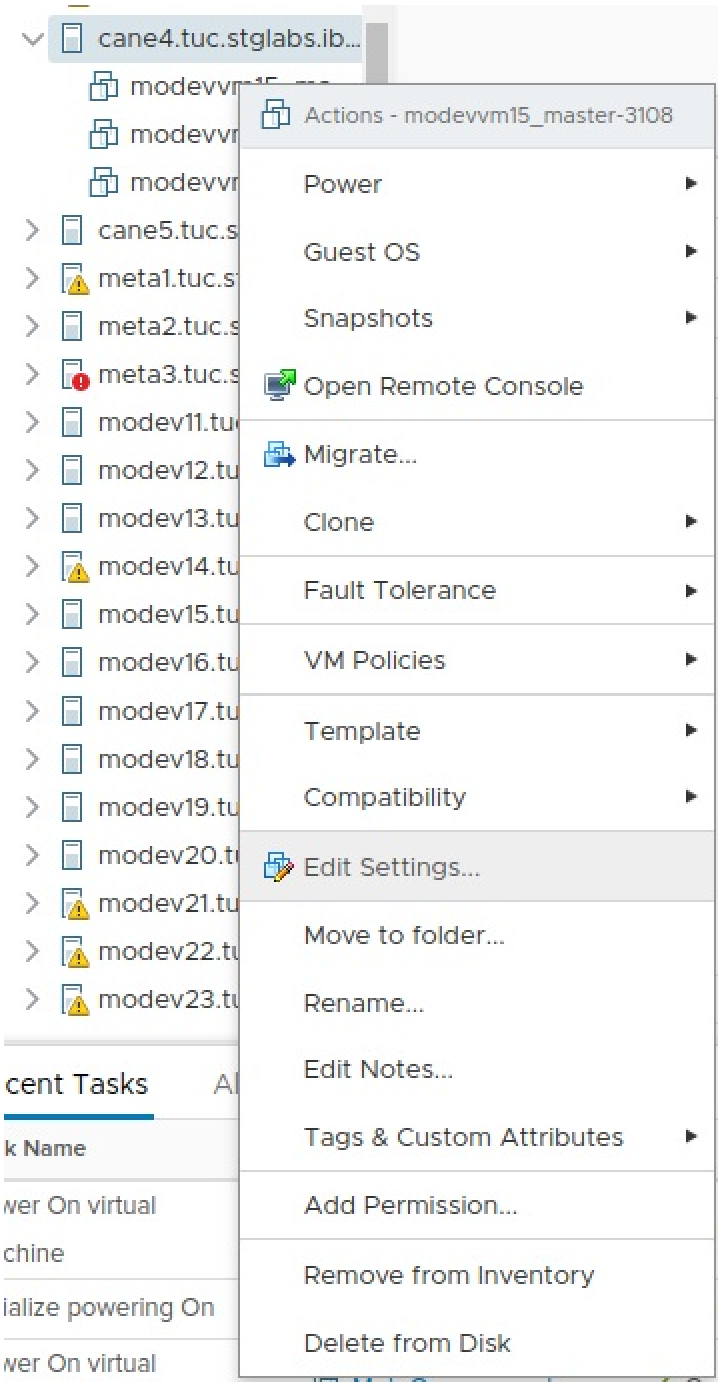
You can use the **VMware vSphere Client** to add the virtual disk required for IBM Spectrum Discover persistent message queues to the virtual appliance.

Before you begin

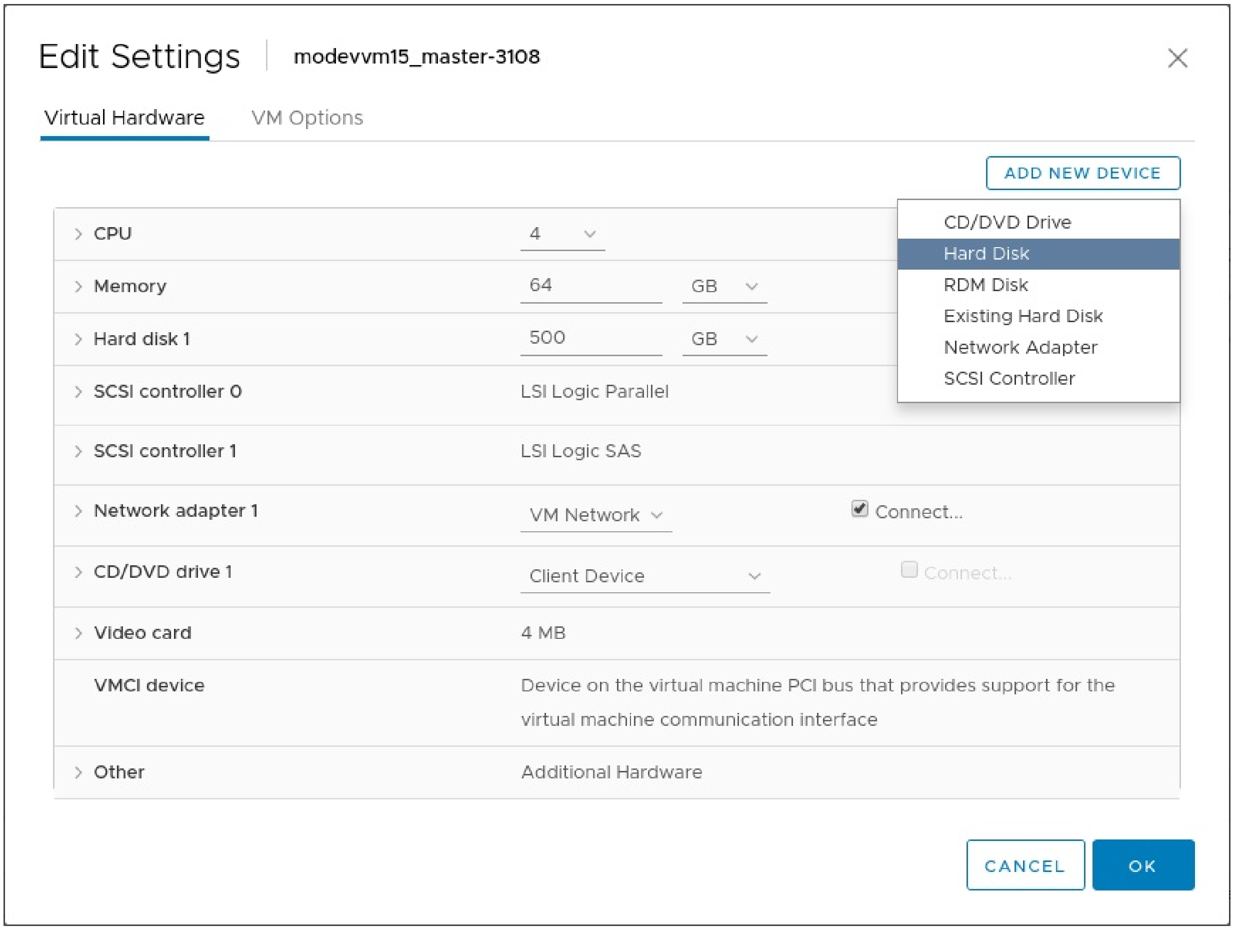
Important: See the section for *Planning* in *IBM Spectrum Discover: Concepts, Planning, and Deployment Guide* for detailed requirements for the persistent message queue VMDK for multi-node deployments. A 3 TB thick provisioned, lazy zeroed VMDK is required for each node in the IBM Spectrum Discover virtual appliance cluster. If an optional IBM Spectrum Discover action agent is to be configured, an additional 500 GB of capacity per node is required.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**.



2. From the **ADD NEW DEVICE** list at the bottom of the dialog box, select **Hard Disk**.

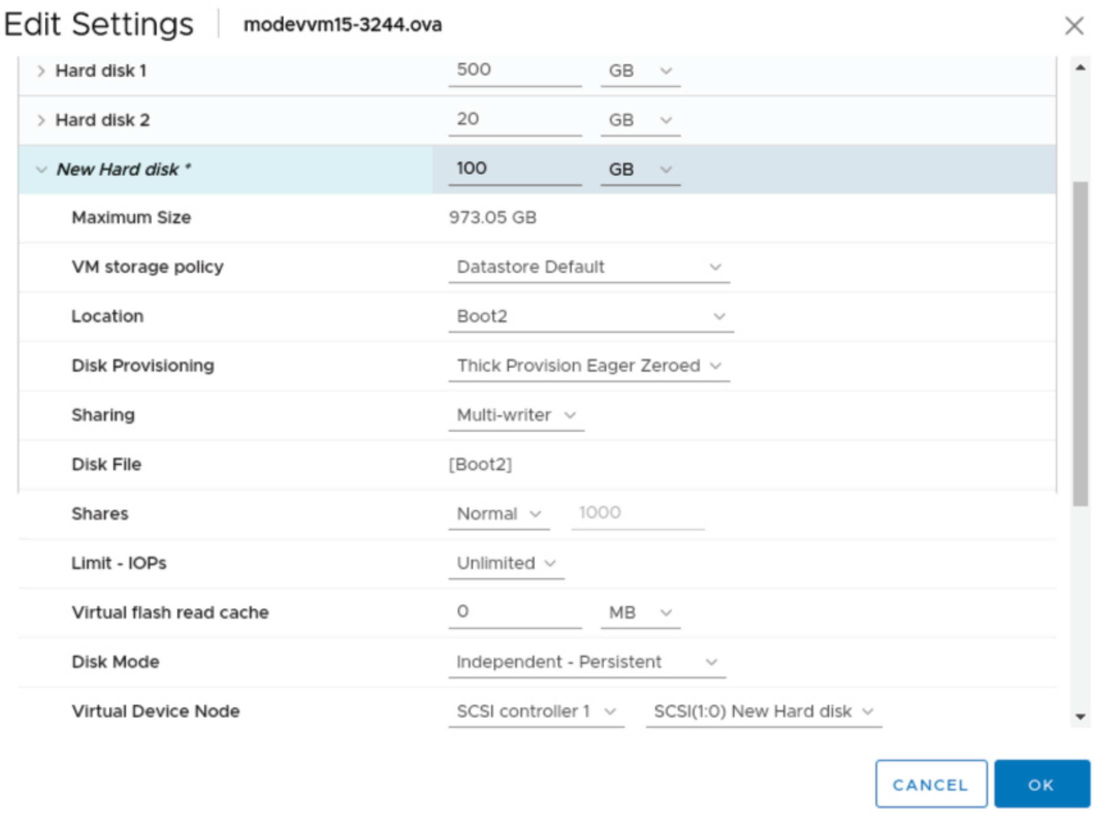


Edit Settings | modevvm15_master-3108

CPU	4	
Memory	64	GB
Hard disk 1	500	GB
SCSI controller 0	LSI Logic Parallel	
SCSI controller 1	LSI Logic SAS	
Network adapter 1	VM Network	☑ Connect...
CD/DVD drive 1	Client Device	□ Connect...
Video card	4 MB	
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface	
Other	Additional Hardware	

A **New Hard Disk** entry appears under **Virtual Hardware**.

3. Click **New Hard Disk** to expand the menu and select options for the disk.



Edit Settings | modevvm15-3244.ova

Hard disk 1	500	GB
Hard disk 2	20	GB
*New Hard disk **	100	GB
Maximum Size	973.05 GB	
VM storage policy	Datastore Default	
Location	Boot2	
Disk Provisioning	Thick Provision Eager Zeroed	
Sharing	Multi-writer	
Disk File	[Boot2]	
Shares	Normal	1000
Limit - IOPs	Unlimited	
Virtual flash read cache	0	MB
Disk Mode	Independent - Persistent	
Virtual Device Node	SCSI controller 1	SCSI(1:0) New Hard disk

pov00025

At this point, you can set the size, provisioning and location of the virtual disk. The default location is the datastore where the virtual appliance resides. But you can select a different datastore if needed.

Note: The above image shows an example of a new hard disk size 20 GB. In practice this number should be much larger. For production environments, it is required to allocate more space for the persistent message queue. See the section for *Planning* in *IBM Spectrum Discover: Concepts, Planning, and Deployment Guide*

4. Click **OK** to confirm your settings and create the virtual disk.

Adding an LSI Logic SCSI Controller to the first virtual appliance node on the IBM Spectrum Discover cluster

You can use the VMware vSphere Client to add a SCSI controller that manages the virtual disks required for IBM Spectrum Discover database to the virtual appliance.

About this task

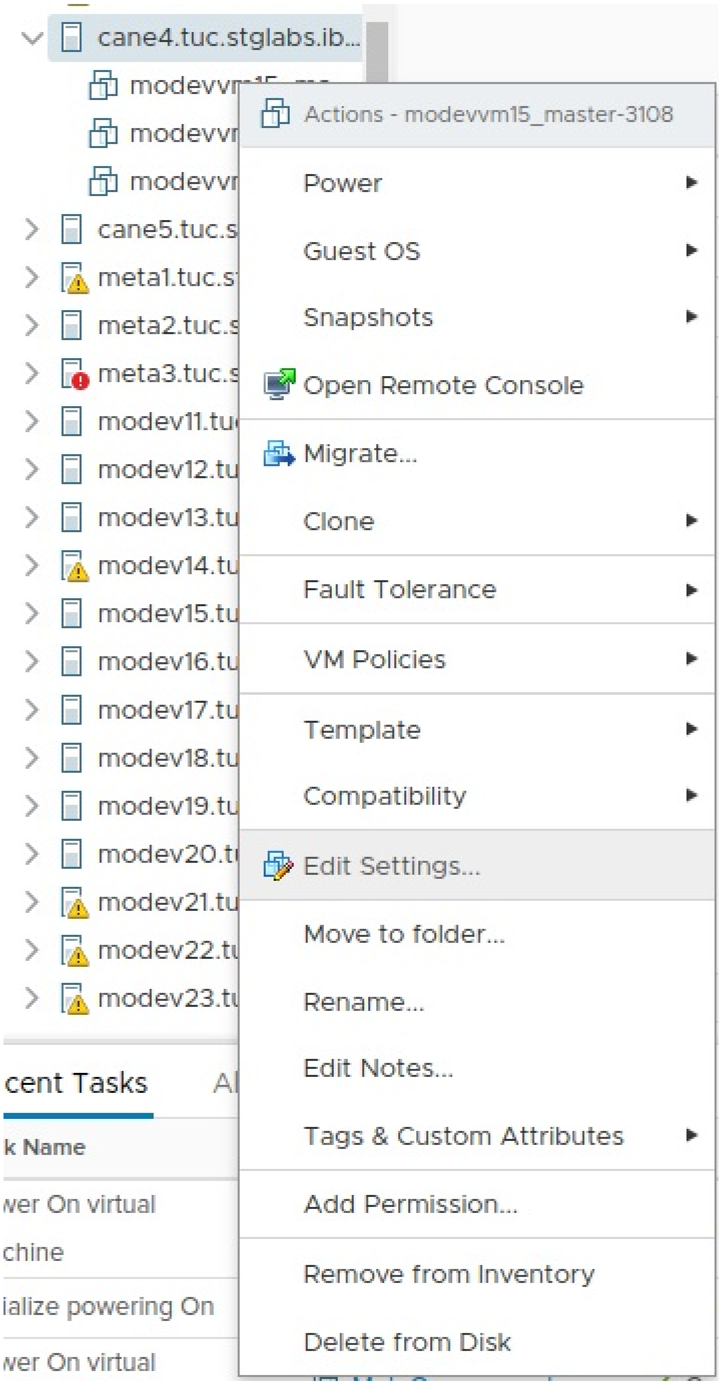
SCSI Controller 0 which is defined as `LSI Logic Parallel` manages the boot and OS VMDK as well as the persistent message queue. SCSI controller 1 must be added to manage the shared virtual disks used for the IBM Spectrum Discover database.

SCSI controller 1 must be set to `LSI Logic` SAS. The SCSI bus sharing mode must be set to either virtual or physical.

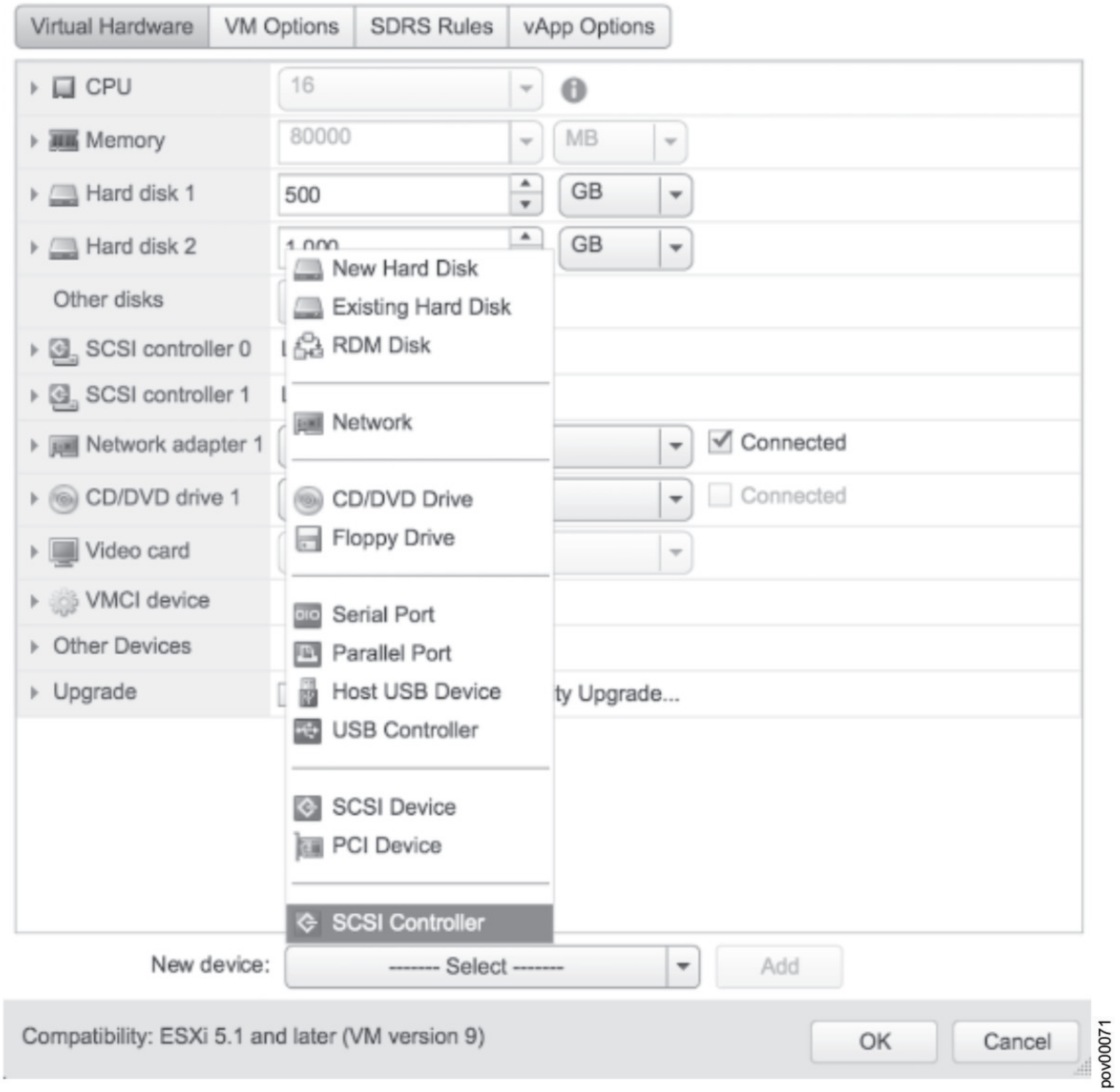
- If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on separate ESXi servers, the SCSI bus sharing mode must be set to `Physical` so that the virtual disks for the database can be shared between virtual machines on any server.
- If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on the same ESXi server, the SCSI bus sharing mode must be set to `Virtual` so that the virtual disks for the database can be shared between the virtual machines on the same server.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**:



2. From the **ADD NEW DEVICE** list at the bottom of the dialog box, select **SCSI Controller**.


CPU: 16

Memory: 80000 MB

Hard disk 1: 500 GB

Hard disk 2: 1,000 GB

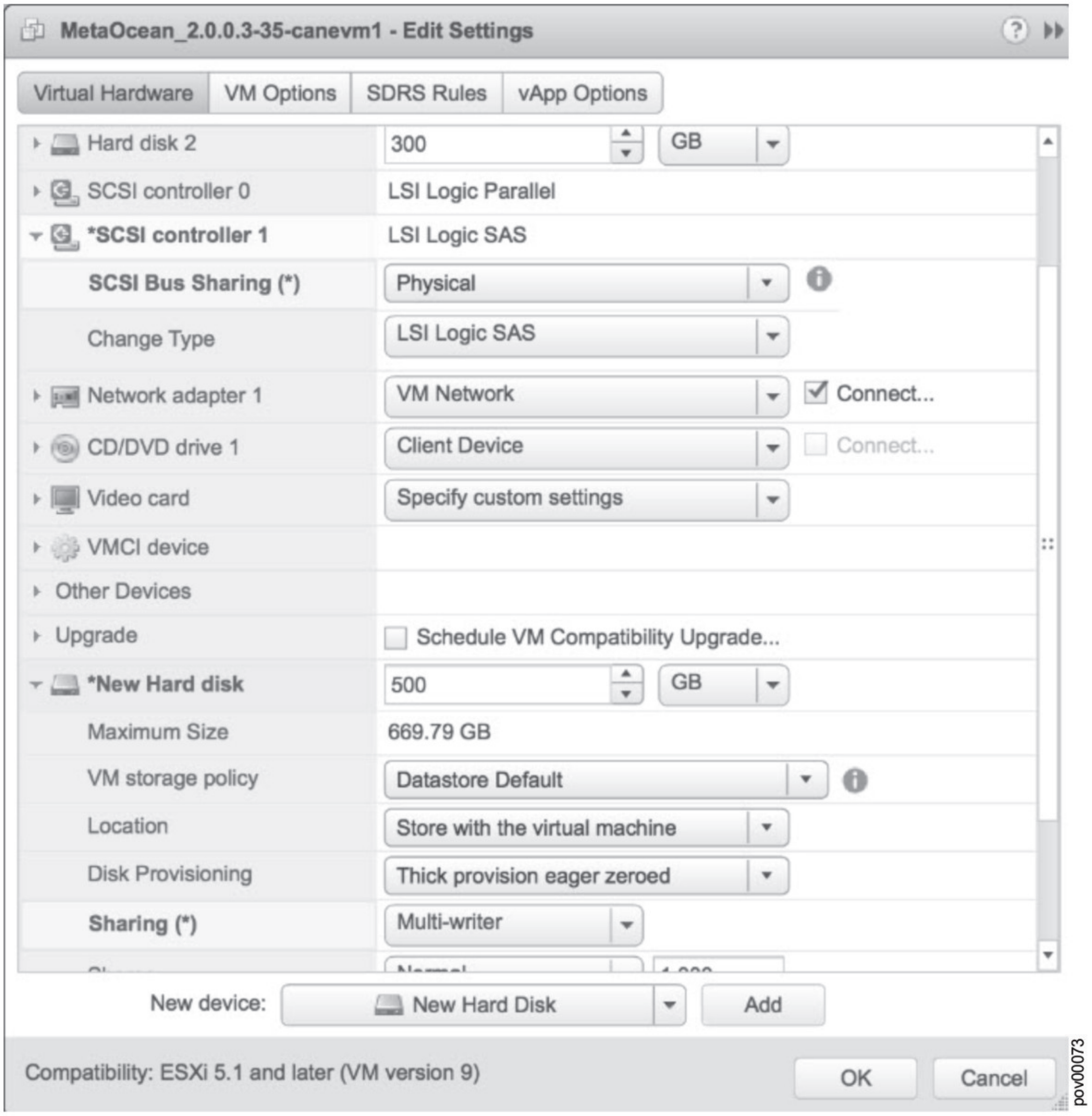
Network adapter 1: ☑ Connected

CD/DVD drive 1: □ Connected

New device: ------- Select -------

Compatibility: ESXi 5.1 and later (VM version 9)

3. Set the SCSI controller type to `LSI Logic SAS`.



MetaOcean_2.0.0.3-35-canevm1 - Edit Settings

Hard disk 2	300 GB	
SCSI controller 0	LSI Logic Parallel	
*SCSI controller 1	LSI Logic SAS	
SCSI Bus Sharing (*)	Physical	
Change Type	LSI Logic SAS	
Network adapter 1	VM Network	☑ Connect...
CD/DVD drive 1	Client Device	□ Connect...
Video card	Specify custom settings	
VMCI device		
Other Devices		
Upgrade	□ Schedule VM Compatibility Upgrade...	
*New Hard disk	500 GB	
Maximum Size	669.79 GB	
VM storage policy	Datastore Default	
Location	Store with the virtual machine	
Disk Provisioning	Thick provision eager zeroed	
Sharing (*)	Multi-writer	

New device: New Hard Disk

Compatibility: ESXi 5.1 and later (VM version 9)

4. Set the SCSI Bus Sharing mode to either `Physical` or `Virtual` based on your IBM Spectrum Discover cluster configuration.

   Note:

   - If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on separate ESXi servers, the SCSI sharing mode must be set to `Physical` so that the virtual disks for the database can be shared between virtual machines on any server.
   - If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on the same ESXi server, the SCSI bus sharing mode must be set to `Virtual` so that the virtual disks for the database can be shared between the virtual machines on the same server.

Adding virtual shared disks for the database to the first node in the IBM Spectrum Discover cluster. You can use the VMware vSphere Client to add the virtual disks required for IBM Spectrum Discover database to the virtual appliance.

Before you begin

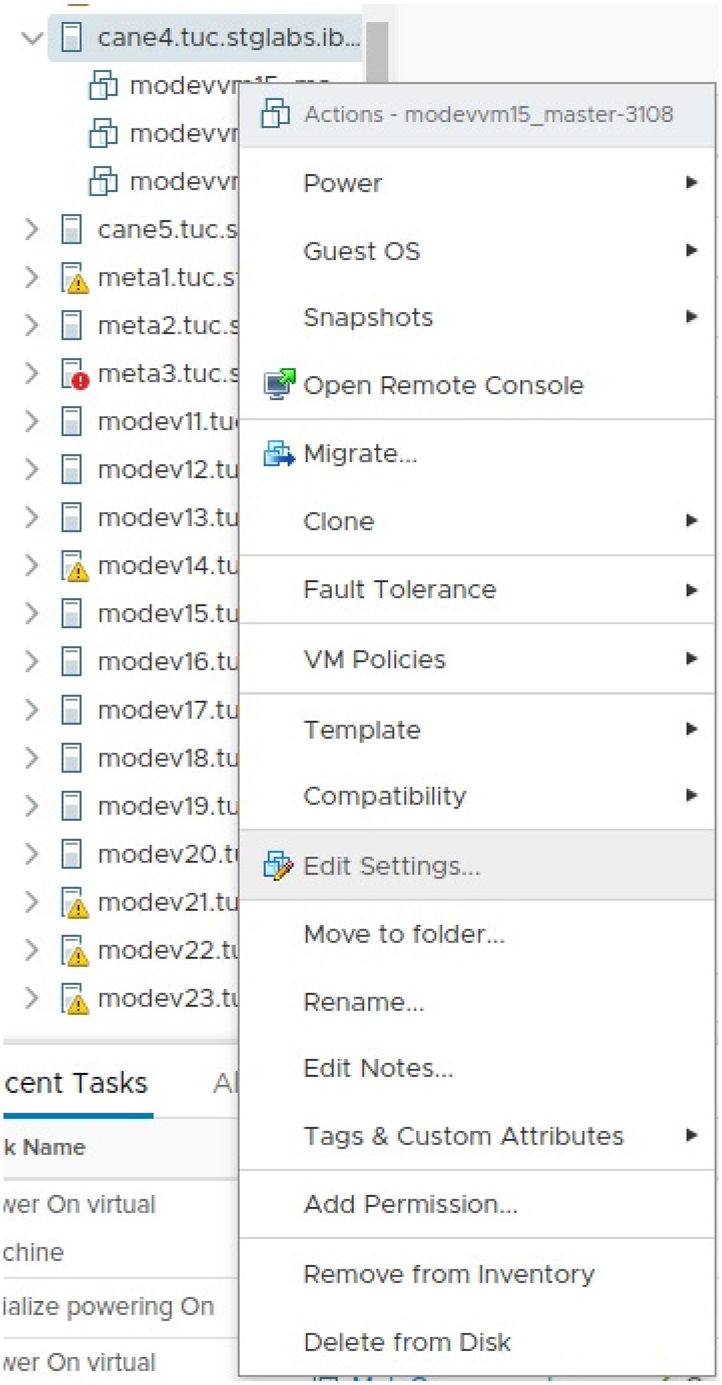
Important: See the section for Planning for detailed requirements for the database VMDK for a multi-node IBM Spectrum Discover virtual appliance cluster. For the database shared VMDK storage device, 14

TB is required to index up to 10 billion files and objects. 2 GB per 2 million files can be used as a capacity sizing metric. The VMDK storage device must be thick provisioned, and eager zeroed in order to be shared between the three nodes in the IBM Spectrum Discover cluster.

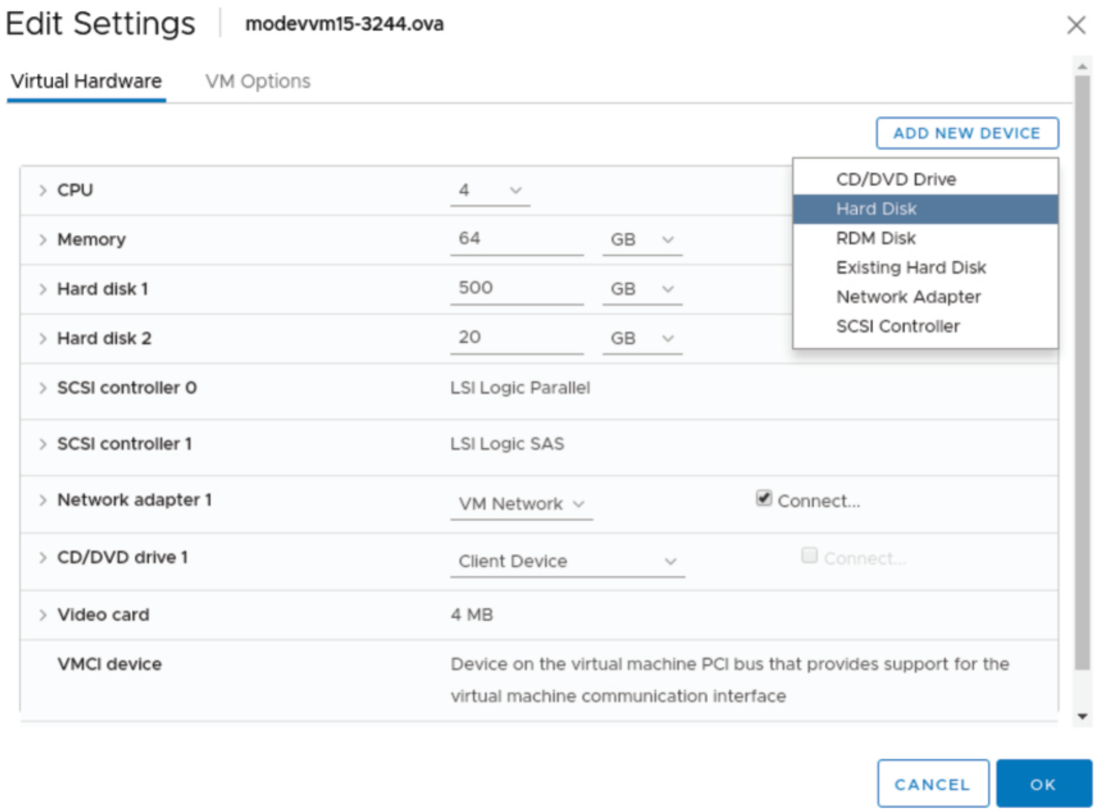
Each VMDK storage device must also be configured to be managed by SCSI Controller 1.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**.



2. From the **ADD NEW DEVICE** list, select `Hard Disk`



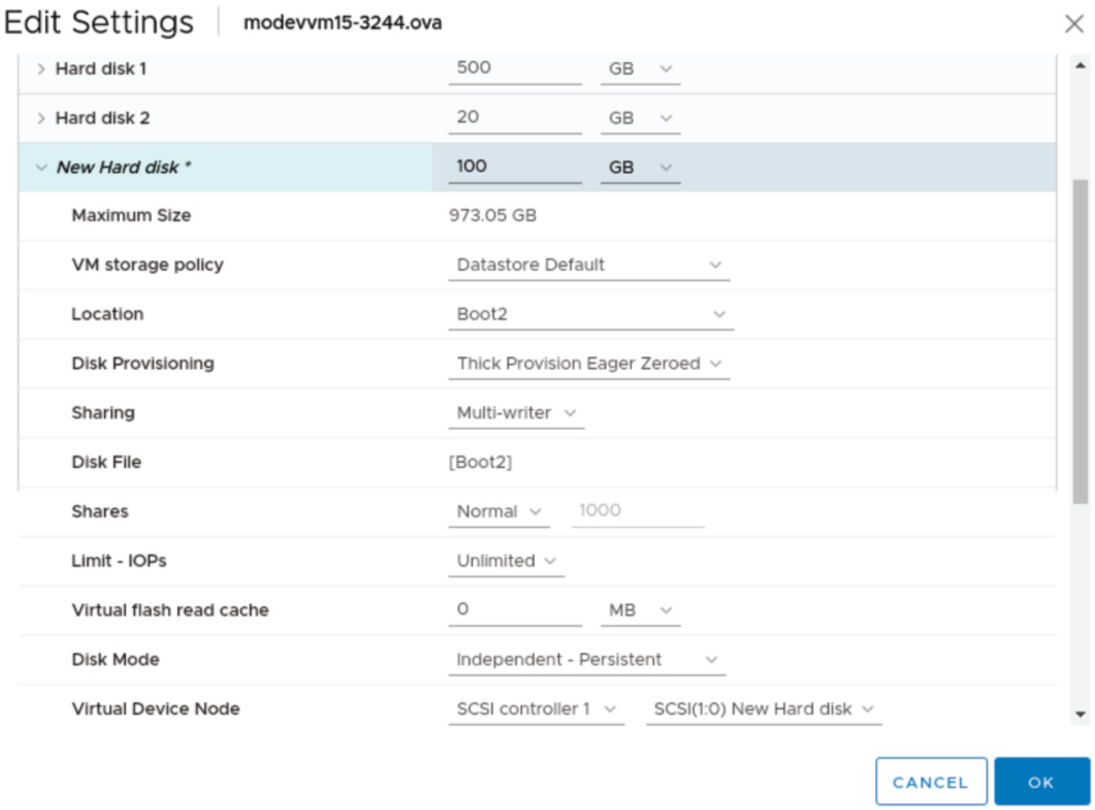
Edit Settings | modevvm15-3244.ova

Virtual Hardware | VM Options

CPU	4	
Memory	64 GB	
Hard disk 1	500 GB	
Hard disk 2	20 GB	
SCSI controller 0	LSI Logic Parallel	
SCSI controller 1	LSI Logic SAS	
Network adapter 1	VM Network	☑ Connect...
CD/DVD drive 1	Client Device	□ Connect...
Video card	4 MB	
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface	

pov00024

3. Click **New Hard Disk** to expand the menu and select the options for the disk.



Edit Settings | modevvm15-3244.ova

Hard disk 1	500 GB	
Hard disk 2	20 GB	
*New Hard disk **	100 GB	
Maximum Size	973.05 GB	
VM storage policy	Datastore Default	
Location	Boot2	
Disk Provisioning	Thick Provision Eager Zeroed	
Sharing	Multi-writer	
Disk File	[Boot2]	
Shares	Normal	1000
Limit - IOPs	Unlimited	
Virtual flash read cache	0 MB	
Disk Mode	Independent - Persistent	
Virtual Device Node	SCSI controller 1	SCSI(1:0) New Hard disk

pov00025

Note:

- **Disk Provisioning** must be set to `Thick Provision Eager Zeroed`.
- **Sharing Mode** must be set to `Multi-writer`
- **Virtual Device** must be set to `SCSI controller 1` and `SCSI (1,0) New Hard disk`
- At this point, you can set the size, provisioning and location of the virtual disk. The default location is the data store where the virtual appliance resides. But you can select a different datastore if needed.

4. Click **OK** to confirm your settings and create the virtual disk.
5. Repeat this procedure to add a total of three shared virtual devices for the IBM Spectrum Discover database.

Adding a virtual disk for IBM Spectrum Discover persistent message queues for the second and third nodes in the multi-node cluster

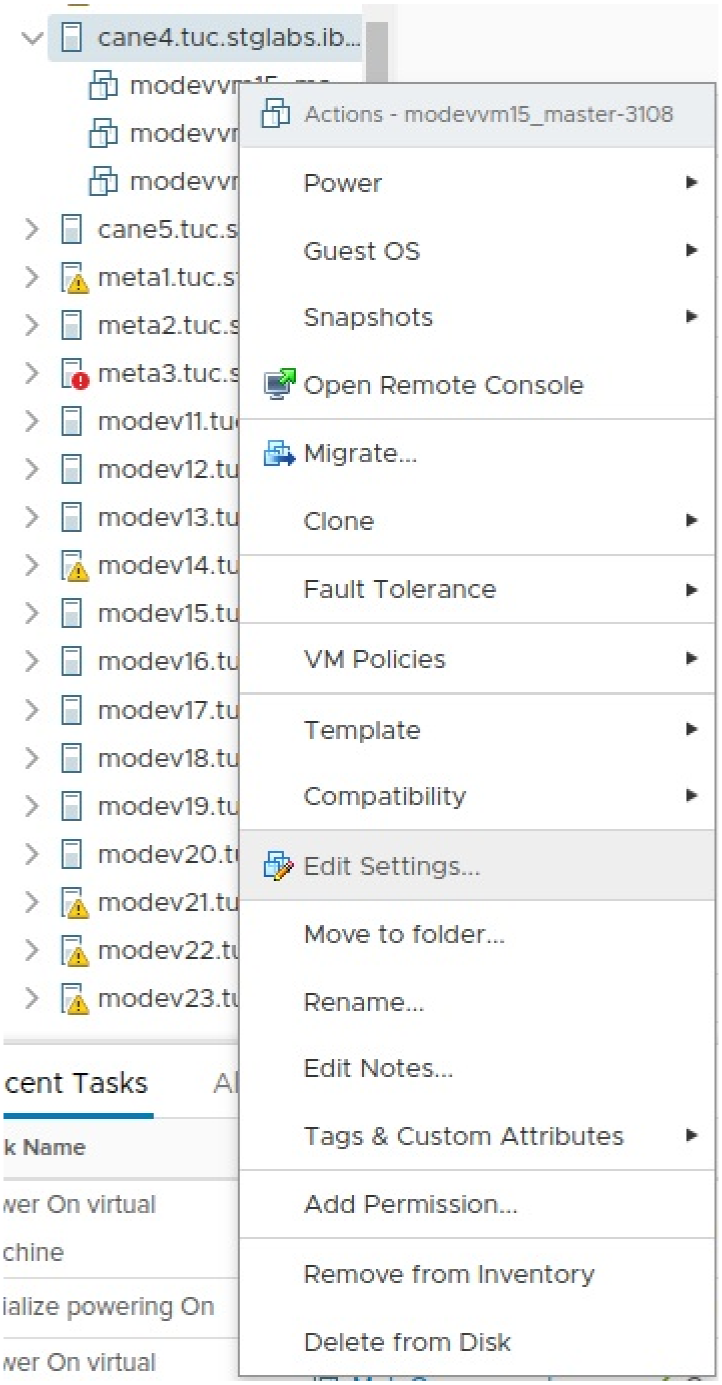
You can use the VMware vSphere Client to add the virtual disk required for IBM Spectrum Discover persistent message queues to the virtual appliance.

Before you begin

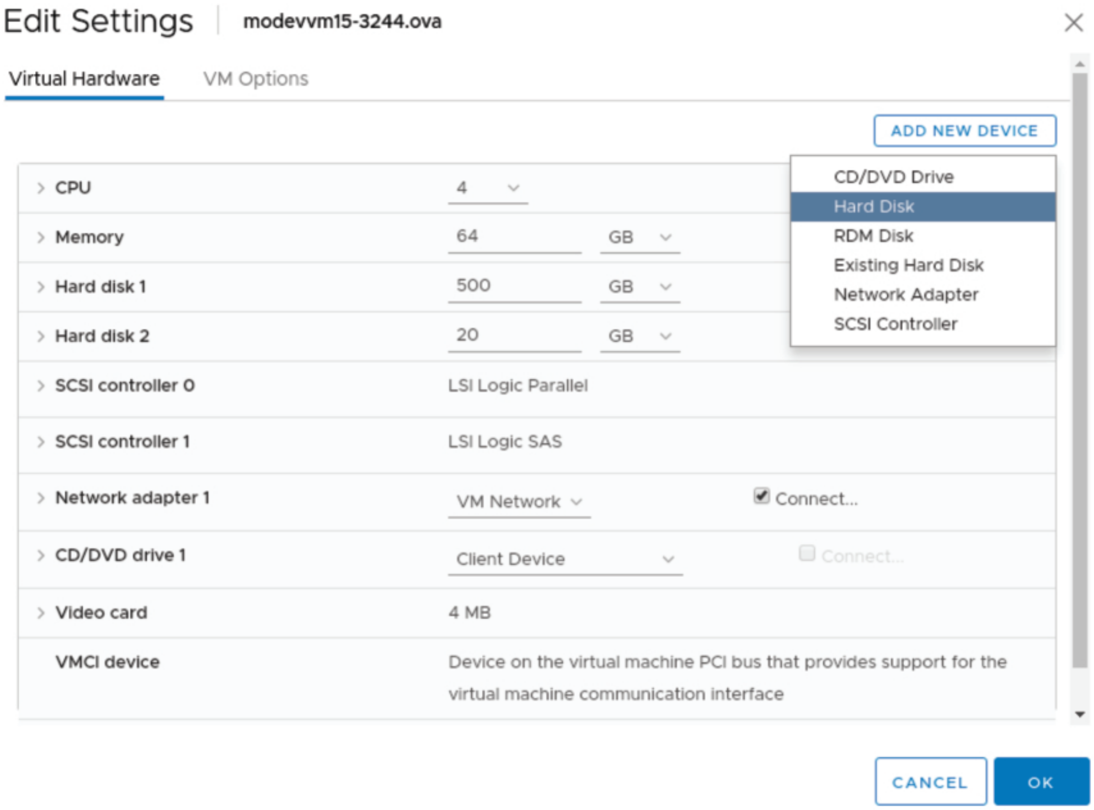
Important: See the section for Planning for detailed requirements for the persistent message queue VMDK for multi-node deployments. A 7 TB thick provisioned, and lazy zeroed VMDK is required for each node in the IBM Spectrum Discover virtual appliance cluster. If an optional IBM Spectrum Discover action agent is to be configured, an additional 1 TB of capacity per node is required.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**.



2. From the **ADD NEW DEVICE** list, select `Hard Disk`



Edit Settings | modevvm15-3244.ova

CPU	4		
Memory	64	GB	
Hard disk 1	500	GB	
Hard disk 2	20	GB	
SCSI controller 0	LSI Logic Parallel		
SCSI controller 1	LSI Logic SAS		
Network adapter 1	VM Network		☑ Connect...
CD/DVD drive 1	Client Device		□ Connect...
Video card	4 MB		
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface		

pov00024

A **New Hard Disk** entry appears under **Virtual Hardware**.

3. Click **New Hard Disk** to expand the menu and select options for the disk.

   At this point, you can set the size, provisioning and location of the virtual disk. The default location is the datastore where the virtual appliance resides. But you can select a different datastore if needed.

   Note: The image in step "2" on page 64 shows an example of a new hard size of 20 GB. In practice this number should be much larger. For production environments, it is required to allocate more space for the persistent message queue. See the section for Chapter 2, "Planning," on page 13.

4. Click **OK** to confirm your settings and create the virtual disk.
5. Repeat steps 1-4 on the third IBM Spectrum Discover node.

Adding LSI Logic SCSI Controller to the second and third virtual appliance nodes in the IBM Spectrum Discover cluster

You can use the VMware vSphere Client to add a SCSI controller that manages the virtual disks required for IBM Spectrum Discover database to the virtual appliance.

Before you begin

Important: SCSI Controller 0 which is defined as LSI Logic Parallel manages the boot and OS VMDK as well as the persistent message queue. SCSI controller 1 must be added to manage the virtual disks used for the IBM Spectrum Discover database.

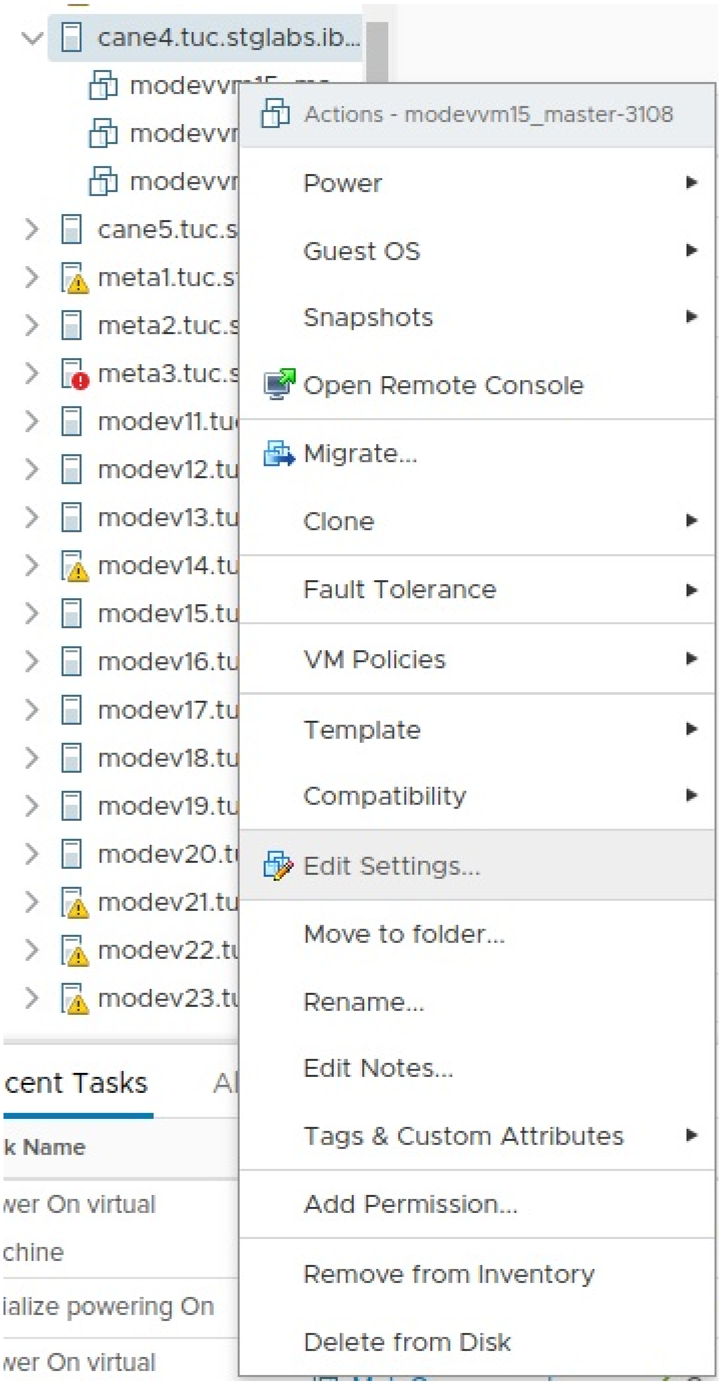
SCSI controller 1 must be set to LSI Logic SAS. The SCSI bus sharing mode must be set to either virtual or physical.

- If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on separate ESXi servers, the SCSI bus sharing mode must be set to `Physical` so that the virtual disks for the database can be shared between virtual machines on any server.

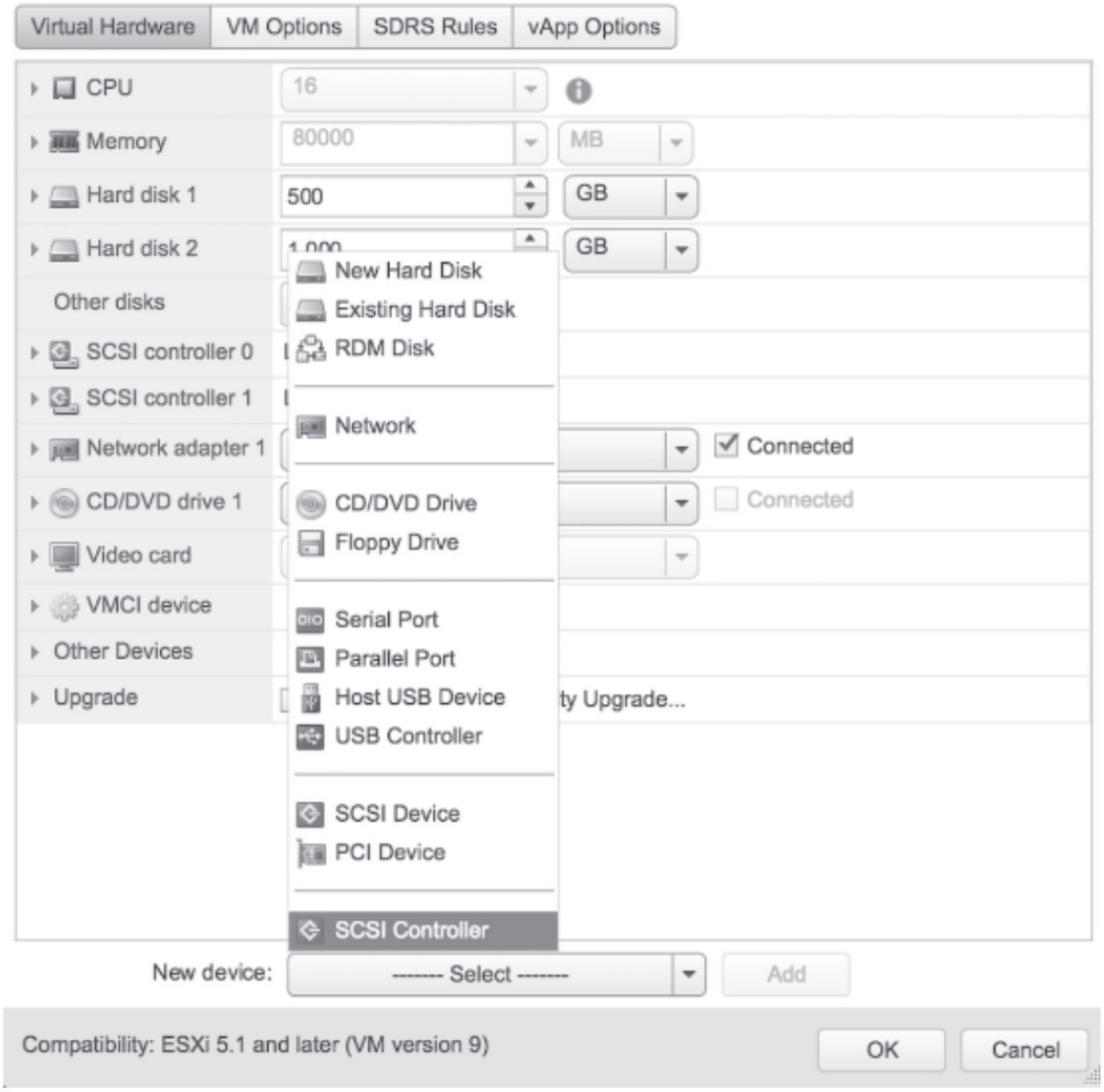
- If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on the same ESXi server, the SCSI bus sharing mode must be set to `Virtual` so that the virtual disks for the database can be shared between the virtual machines on the same server.

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**.



2. From the **ADD NEW DEVICE** list at the bottom of the dialog box, select **SCSI Controller**.



CPU: 16

Memory: 80000 MB

Hard disk 1: 500 GB

Hard disk 2: 1,000 GB

Network adapter 1: ☑ Connected

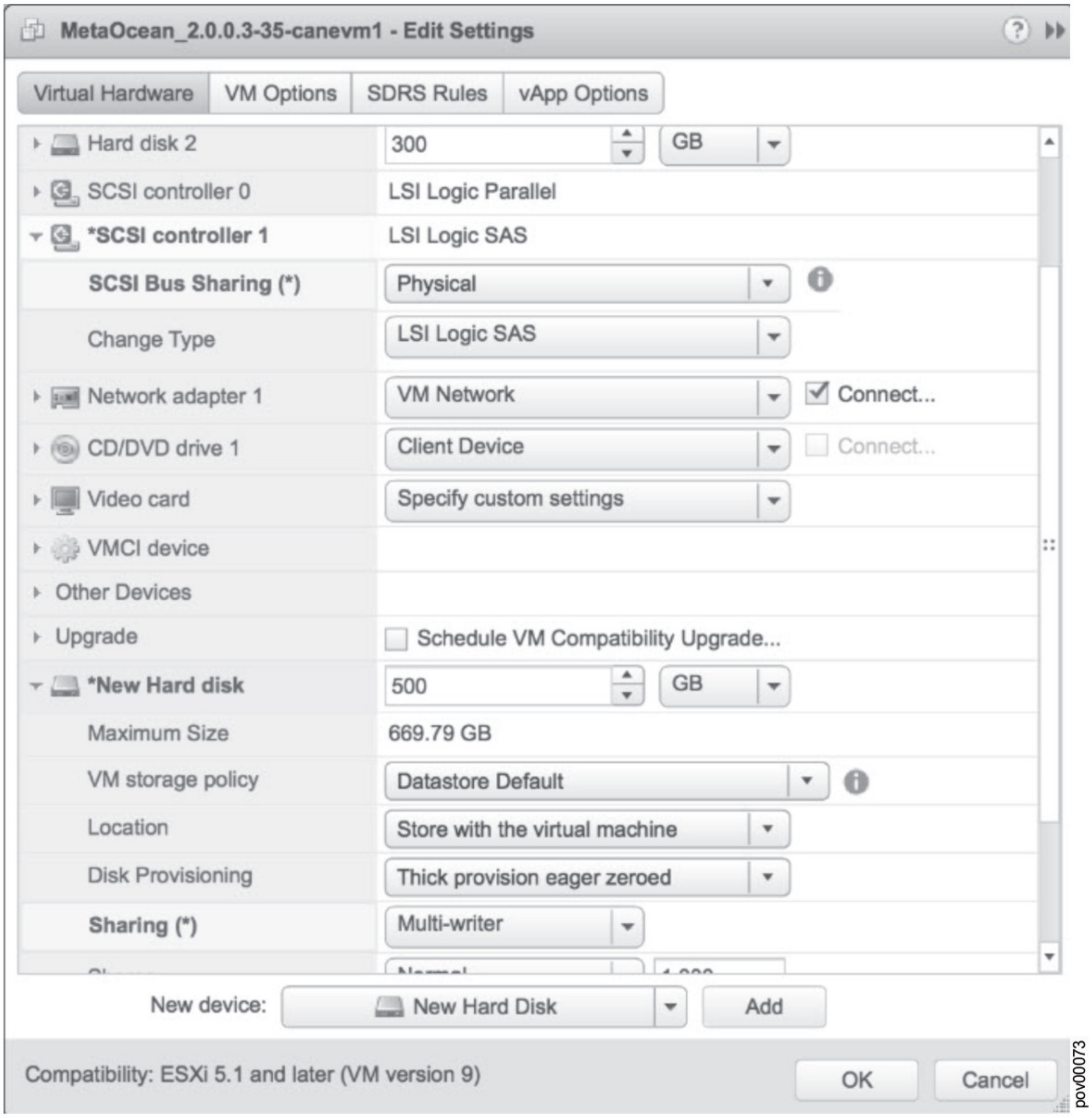
CD/DVD drive 1: □ Connected

New device: SCSI Controller

Compatibility: ESXi 5.1 and later (VM version 9)

pov00071

3. Set the SCSI controller type to `LSI Logic SAS`.



MetaOcean_2.0.0.3-35-canevm1 - Edit Settings

Hard disk 2	300 GB
SCSI controller 0	LSI Logic Parallel
*SCSI controller 1	LSI Logic SAS
SCSI Bus Sharing (*)	Physical
Change Type	LSI Logic SAS
Network adapter 1	VM Network ☑ Connect...
CD/DVD drive 1	Client Device □ Connect...
Video card	Specify custom settings
VMCI device	
Other Devices	
Upgrade	□ Schedule VM Compatibility Upgrade...
***New Hard disk**	500 GB
Maximum Size	669.79 GB
VM storage policy	Datastore Default
Location	Store with the virtual machine
Disk Provisioning	Thick provision eager zeroed
Sharing (*)	Multi-writer

Compatibility: ESXi 5.1 and later (VM version 9)

4. Set the **SCSI Bus Sharing** mode to either `Physical` or `Virtual` based on your IBM Spectrum Discover cluster configuration.

   Note:

   - If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on separate ESXi servers, the SCSI bus sharing mode must be set to `Physical` so that the virtual disks for the database can be shared between virtual machines on any server.
   - If the IBM Spectrum Discover virtual machines that make up the three-node cluster reside on the same ESXi server, the SCSI bus sharing mode must be set to `Virtual` so that the virtual disks for the database can be shared between the virtual machines on the same server.

5. Repeat steps 1-4 on the third IBM Spectrum Discover node.

Adding shared virtual disks for the second and third nodes in the IBM Spectrum Discover cluster

You can use the VMware vSphere Client to add the shared virtual disks to the other two IBM Spectrum Discover virtual machines.

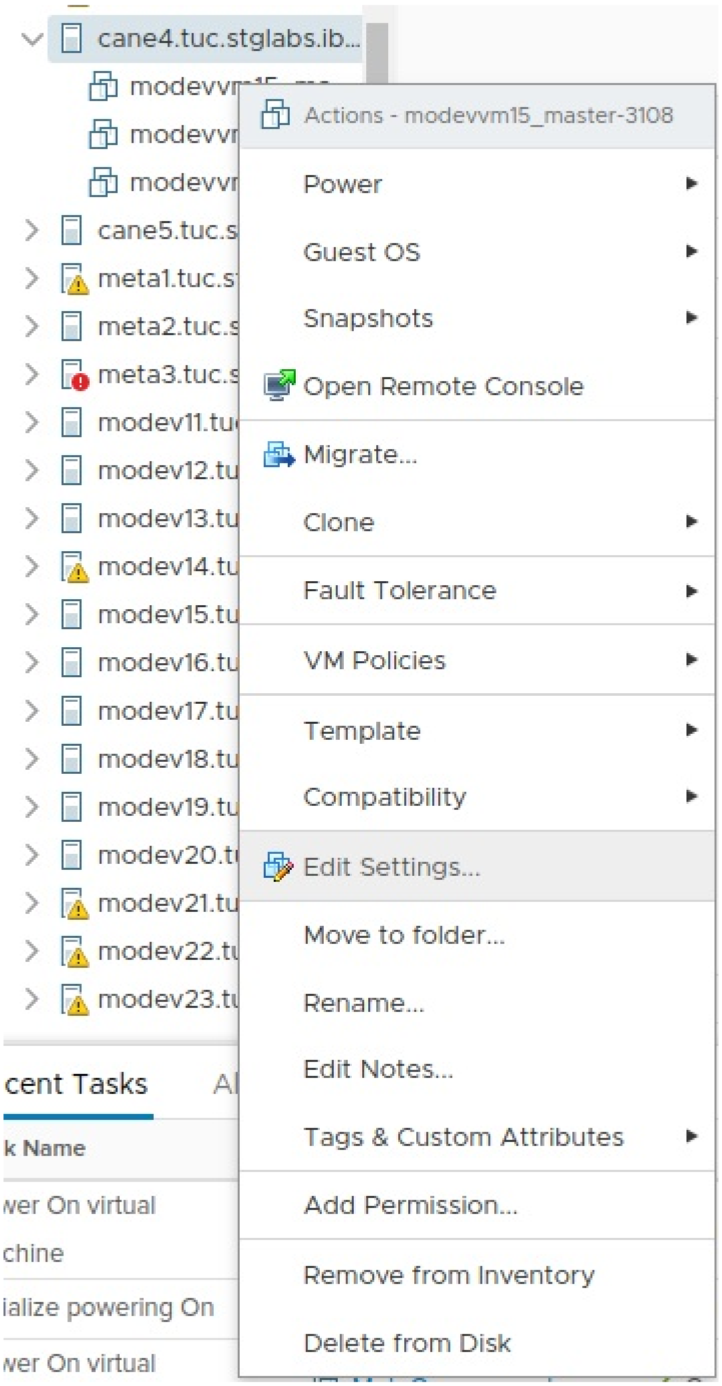
Before you begin

Important: The three shared virtual disks created on the first IBM Spectrum Discover virtual machine must be presented to the second and third IBM Spectrum Discover virtual machines in the three node IBM Spectrum Discover cluster.

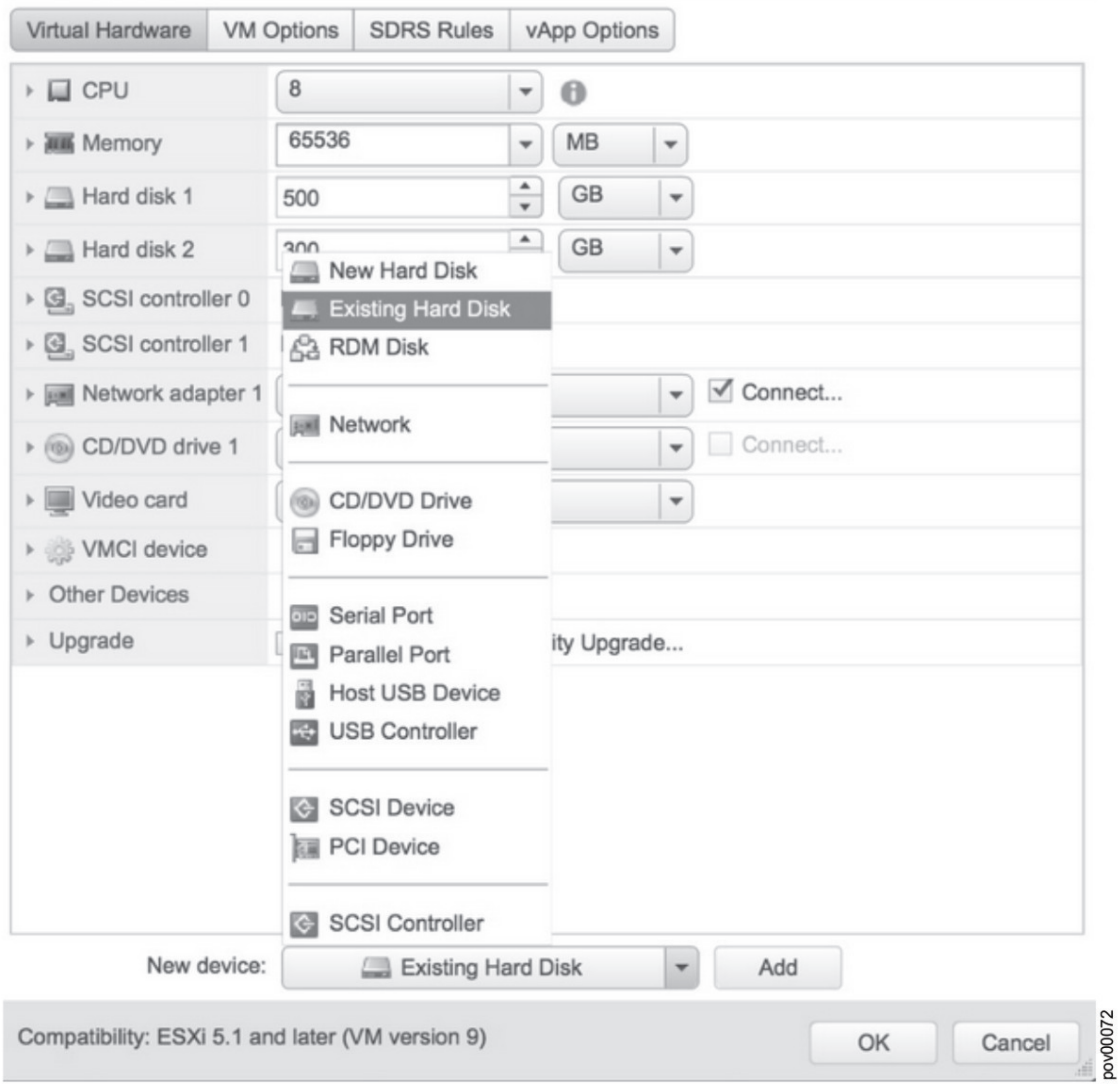
On the virtual machine in the IBM Spectrum Discover cluster perform the following:

Procedure

1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance to which you want to add the virtual disk and click **Edit Settings**.



2. From the **ADD NEW DEVICE** list, select `Existing Hard Disk`.



CPU: 8

Memory: 65536 MB

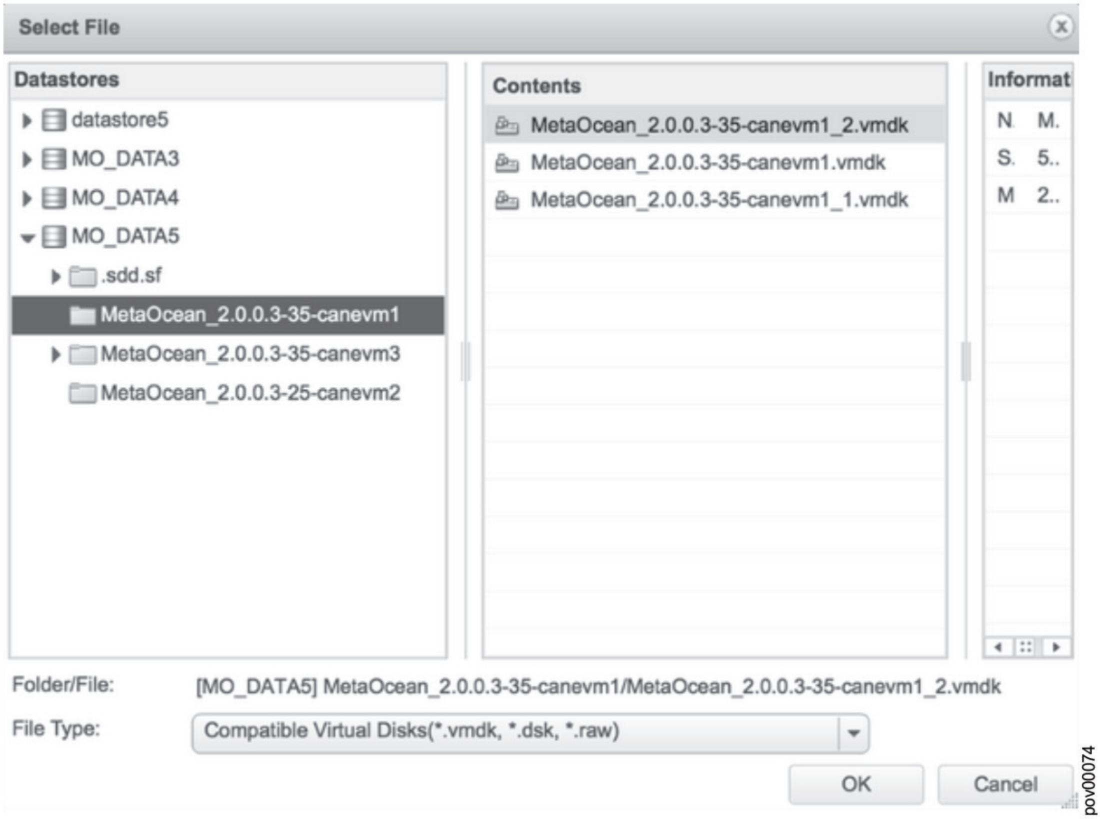
Hard disk 1: 500 GB

Hard disk 2: 300 GB

New device: Existing Hard Disk

Compatibility: ESXi 5.1 and later (VM version 9)

3. Navigate to the datastore containing the first shared disk and select the shared VMDK device. Then click **OK**.


Select File
Datastores
datastore5
MO_DATA3
MO_DATA4
MO_DATA5
.sdd.sf
MetaOcean_2.0.0.3-35-canevm1
MetaOcean_2.0.0.3-35-canevm3
MetaOcean_2.0.0.3-25-canevm2
Contents
MetaOcean_2.0.0.3-35-canevm1_2.vmdk
MetaOcean_2.0.0.3-35-canevm1.vmdk
MetaOcean_2.0.0.3-35-canevm1_1.vmdk
Informat
N. M.
S. 5..
M 2..
OK
Cancel
pov00074

Folder/File: [MO_DATA5] MetaOcean_2.0.0.3-35-canevm1/MetaOcean_2.0.0.3-35-canevm1_2.vmdk

File Type: Compatible Virtual Disks(*.vmdk, *.dsk, *.raw)

4. Ensure that the **Sharing Mode** is set to `multi-writer` and that the **Virtual Device Node** is set to `SCSI Controller` and `SCSI Controller 1`
5. Repeat steps 1-3 for the second and third shared storage virtual disk
6. After completing steps 1-4 on the second IBM Spectrum Discover virtual appliance node, repeat steps 1-4 on the third IBM Spectrum Discover virtual appliance node

Configuring CPU and memory allocation for a multi-node IBM Spectrum Discover virtual appliance cluster

It is a requirement to increase the default allocations of CPU and memory for each IBM Spectrum Discover virtual appliance.

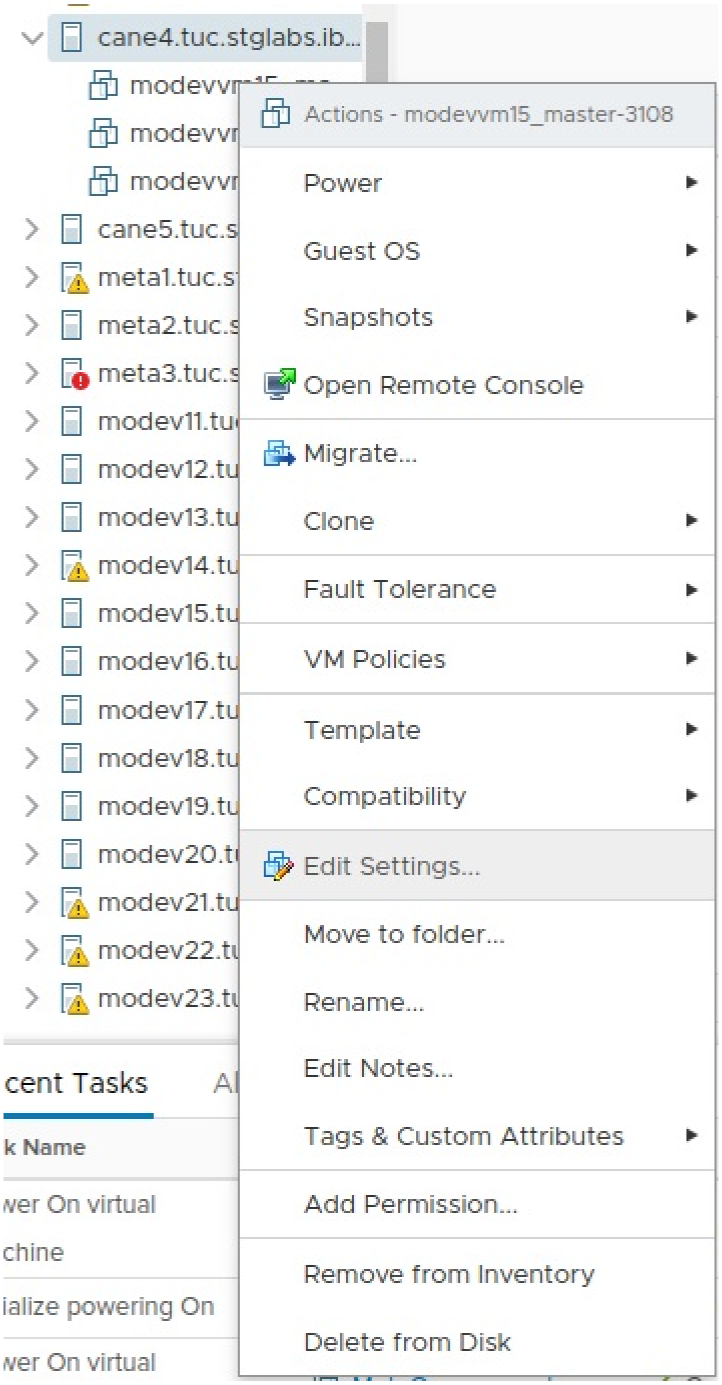
About this task

It is recommended to reserve all the memory assigned to the IBM Spectrum Discover virtual appliance to avoid running out of physical memory and swapping.

Note: Each node on a multi-node production IBM Spectrum Discover virtual appliance cluster requires 256 GB RAM and 32 logical processors. See the section for Planning in the IBM Spectrum Discover Knowledge Center.

Procedure

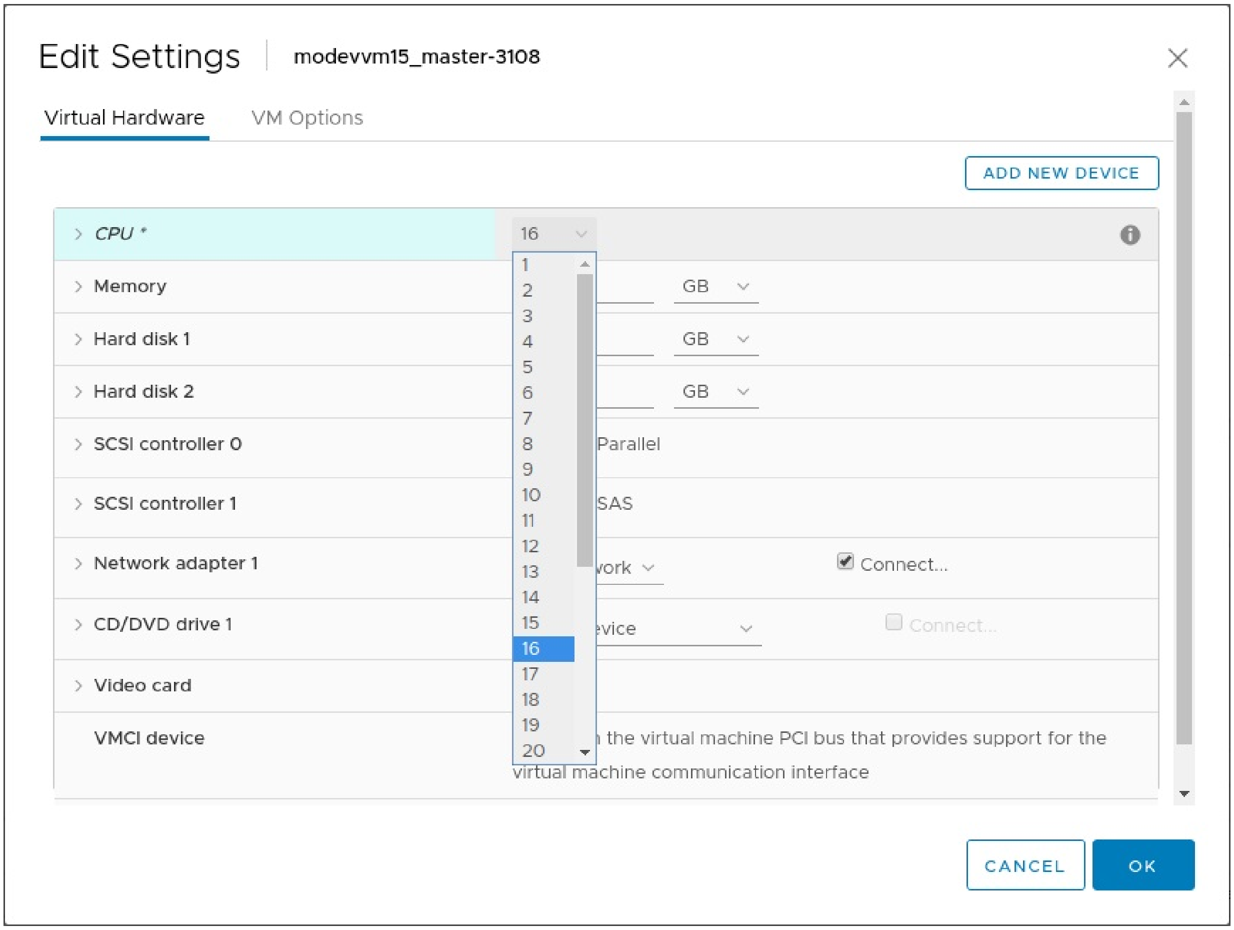
1. In the vSphere Client, right-click the IBM Spectrum Discover virtual appliance for which you want to change the CPU and memory allocation and click **Edit Settings**.



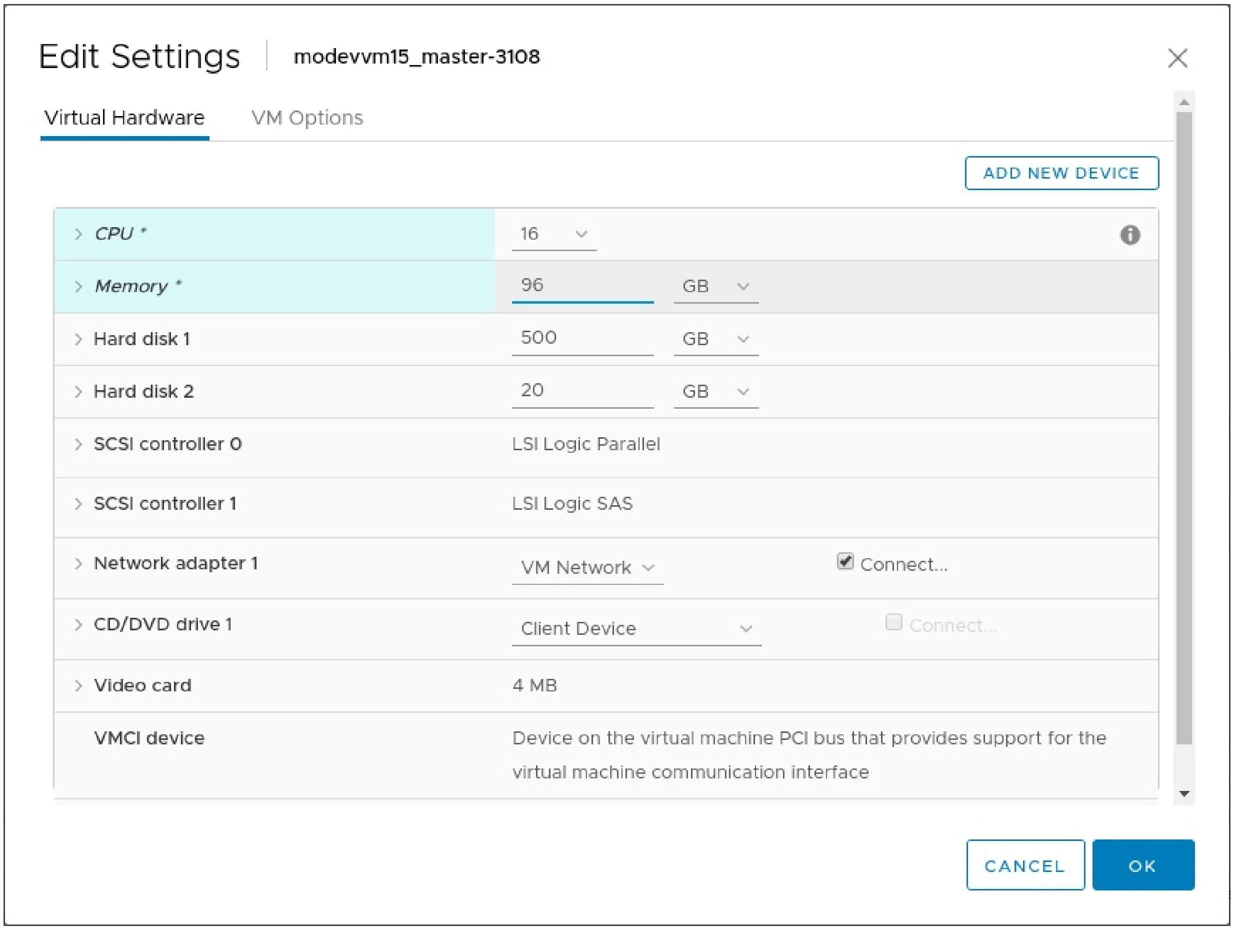
Important

Each node on a multi-node production IBM Spectrum Discover virtual appliance cluster requires 256 GB RAM and 32 logical processors. See the section for Planning.

2. Under **Virtual Hardware**, from the **CPU** list, select the number that you want to set for CPU allocation.



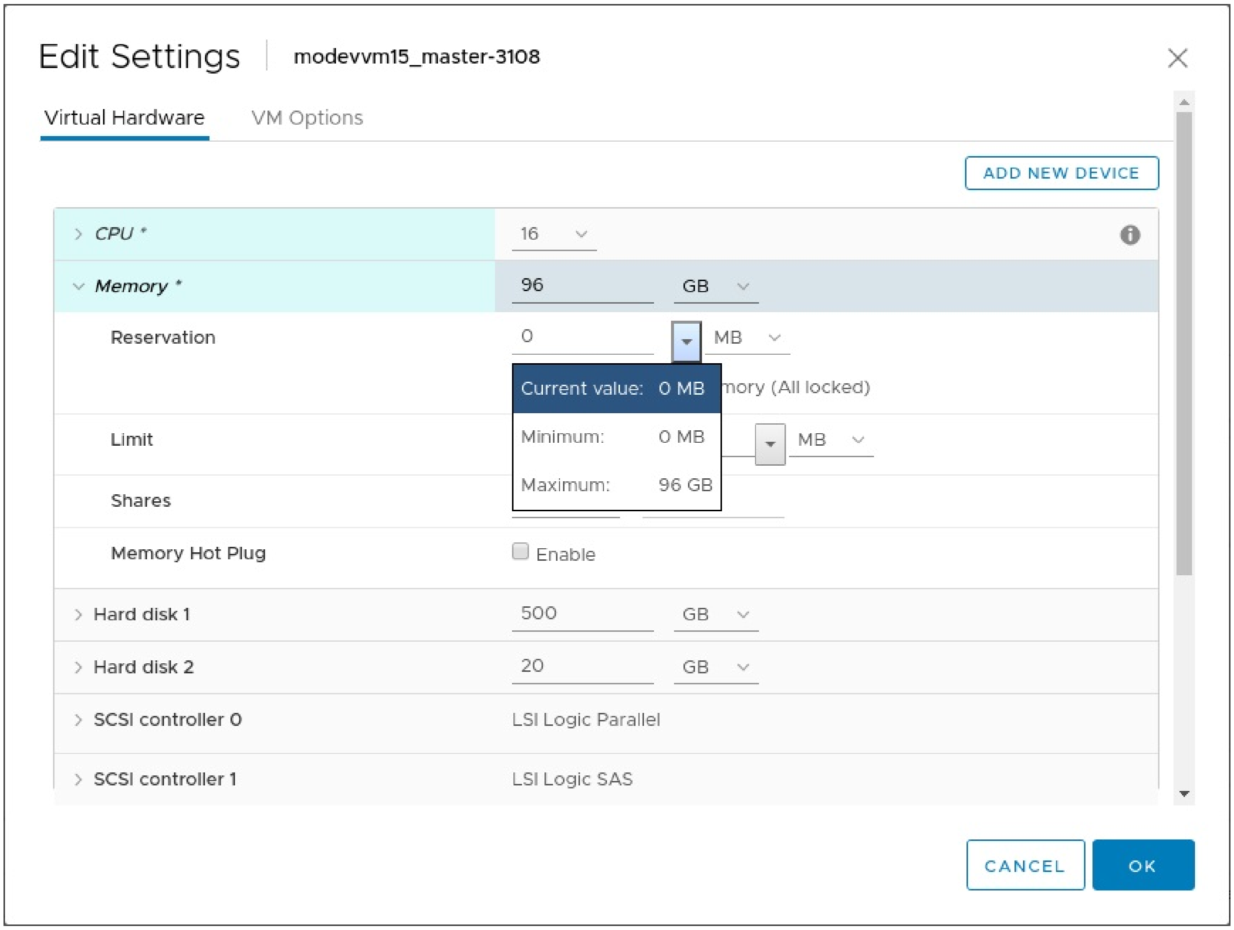
3. In the **Memory** field, enter the number that you want to set for memory allocation and select the memory unit from the drop-down list.



modevvm15_master-3108

CPU *	16		
Memory *	96	GB	
Hard disk 1	500	GB	
Hard disk 2	20	GB	
SCSI controller 0	LSI Logic Parallel		
SCSI controller 1	LSI Logic SAS		
Network adapter 1	VM Network		☑ Connect...
CD/DVD drive 1	Client Device		□ Connect...
Video card	4 MB		
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface		

4. In the **Reservation** field under **Memory**, change the number according to the changed memory allocation and select the memory unit from the drop-down list.


Edit Settings | modevvm15_master-3108

Virtual Hardware | VM Options

ADD NEW DEVICE

CPU * 16

Memory * 96 GB

Reservation 0 MB

Current value: 0 MB

Minimum: 0 MB

Maximum: 96 GB

mory (All locked)

Limit MB

Shares

Memory Hot Plug □ Enable

Hard disk 1 500 GB

Hard disk 2 20 GB

SCSI controller 0 LSI Logic Parallel

SCSI controller 1 LSI Logic SAS

CANCEL OK

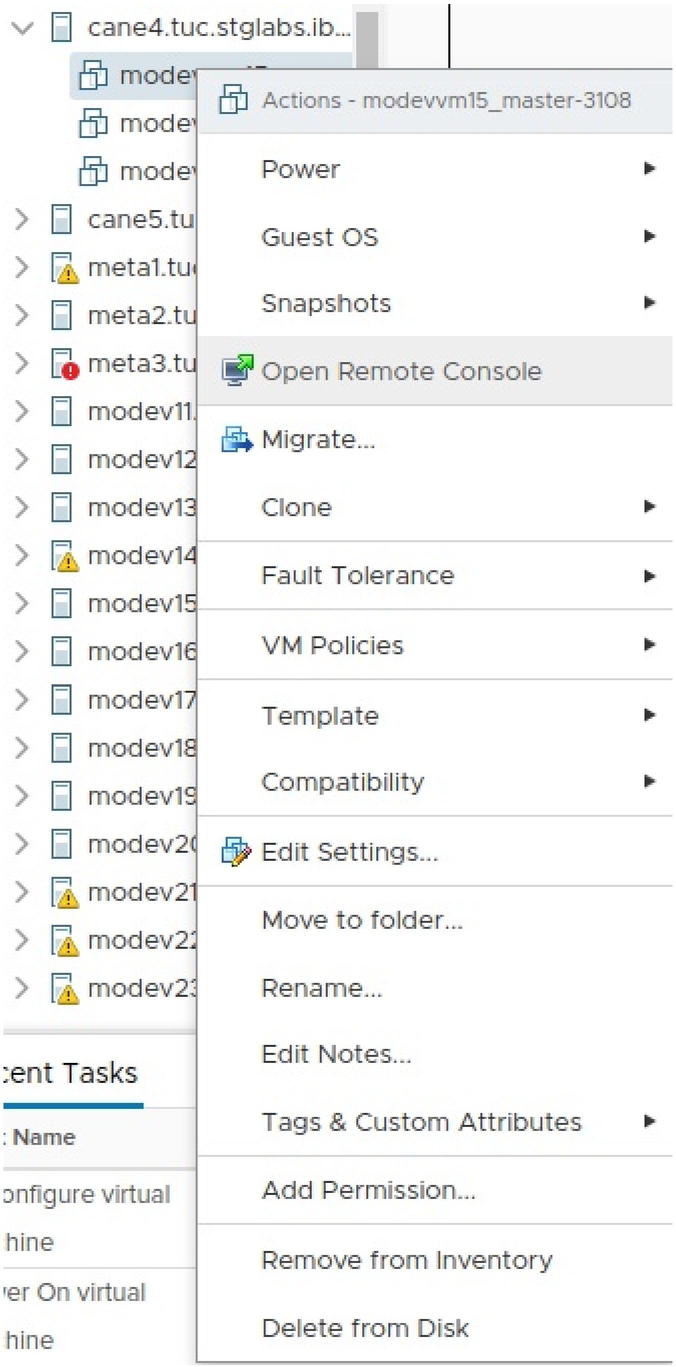
5. Click **OK** to confirm the changes in CPU and memory allocation.

Configuring networking and perform provisioning for the multi-node IBM Spectrum Discover virtual appliance cluster

After virtual appliances in the IBM Spectrum Discover are deployed, and storage, CPU, and memory are configured, you need to configure networking and then provision the virtual appliances by using a provisioning tool. For the multi-node cluster, you need to designate the master node and worker nodes.

Procedure

1. Power on the virtual appliance. In case of a multi-node deployment, power on the virtual appliance that you want to designate as the master node in the IBM Spectrum Discover cluster.
2. In the vSphere Client, right-click the virtual appliance and click **Open Remote Console**.



3. On the virtual appliance login prompt, enter the user name and the password. The default user name is **moadmin** and the default password is **Passw0rd**.
4. Change the directory to `/opt/ibm/metaocean/configuration`.

   ```
   cd /opt/ibm/metaocean/configuration
   ```

   Note: IBM Spectrum Discover requires a fully qualified domain name (FQDN) that is registered in a customer supplied domain name server (DNS). The customer supplied FQDN must be resolvable by the customer supplied DNS from the IBM Spectrum Discover node for the IBM Spectrum Discover virtual appliance to operate properly.
5. Configure the IBM Spectrum Discover virtual appliance networking settings by using the following command:

   ```
   sudo ./mmconfigappliance
   ```

Note: The IBM Spectrum Discover node(s) must be able to communicate with a customer supplied network time protocol (NTP) server in order to operate properly.

It is easier to start with the slave nodes and configure the slave nodes before you configure the master node.

The following table lists the definitions of the required settings:

Parameter	Value format	Recommended value	Example
HostName	host.domain.com	The fully qualified domain name of the node.	node1234.example.co m
Interface	ensXXX	The Ethernet interface to use for the virtual appliance networking.	ens192
IPAddress	xxx.xxx.xxx.xxx	The IP address of the node.	10.10.200.10
NetMask	xxx.xxx.xxx.xxx	The network mask for the IP range of the node.	255.255.254.0
Gateway	host.domain.com	The IP address of the network gateway.	10.10.200.1
DNS	ensXXX	The IP address of a single DNS server.	10.10.200.35
NTPServer	xxx.xxx.xxx.xxx or host.domain.com	The fully qualified domain or the IP address of the NTP server.	pool1.ntp.org
Mode	single or multi	multi	multi
Node type	master or worker	There should be 1 master node and the rest defined as worker nodes	master

Note: [During this step, you are prompted to:

- Set the timezone. If you wish to change it so something other than UTC, you can go through the prompts to choose your continent, country, or region. (No other setup is required to set the timezone.)
- Change the`moadmin` password.

]

[The **`mmconfigappliance`** process takes approximately 1 hour to complete for a production system.]

[Check for the Kubernetes `network_cidr` and `service_cluster_ip_range` values. These specified values must not conflict with the existing host network IP data. The default values are:

```
network_cidr: 10.1.0.0/16
service_cluster_ip_range: 10.0.0.0/16
```

]

[If you are prompted for the `network_cidr` or `service_cluster_ip_range`, consider this information to avoid network conflicts. Private network ranges that might be a good choice are:

- `10.0.0.0` to `10.255.255.255`
- `172.16.0.0` to `172.31.255.255`

]

[For example, you can enter:

- `172.31.0.0/16` for the `network_cidr`
- `172.30.0.0/16` for the `service_cluster_ip_range`

```
network_cidr: 172.31.0.0/16
service_cluster_ip_range: 172.30.0.0/16
```

]

[IBM Spectrum Discover checks to see whether the system host IP overlaps with the Kubernetes default network and service IP range. If this overlap is detected, you get an error message similar to this:

```
Host network (10.1.10.10) is overlapped with the default Kubernetes network (10.1.0.0/16).
Please enter in a new value for the Kubernetes network.
```

]

Upon successful completion, an output similar to the following is displayed.

```
PLAY RECAP ***********************************************************
**********************************************************************
**************************************
203.0.113.14        : ok=241  changed=209  unreachable=0    failed=0
canevm7.example.com : ok=9    changed=6    unreachable=0    failed=0
canevm8.example.com : ok=9    changed=6    unreachable=0    failed=0
canevm9.example.com : ok=9    changed=6    unreachable=0    failed=0
```

The process is completed successfully when you do not see any messages that say `failed` or when you see a message that the failed count = 0.

Known issues with deploying and configuring for multi-node

In case of any errors that you might encounter while deploying or configuring IBM Spectrum Discover, review the following information for details and possible workarounds:

<table>
<tr><th>Issue</th><th>Description</th><th>Resolution or workaround</th></tr>
<tr><td>Log shows error after deployment</td><td>Even after successful deployment, log might show some errors with messages similar to the following:
<pre>
2018-10-15 11:33:00,557 p=12895 u=root |
fatal: [203.0.113.18]:
FAILED! => {"changed": false, "cmd": "awk '/
sse4_2/ {exit 42}'
/proc/cpuinfo", "delta": "0:00:00.004105",
"end":
"2018-10-15 11:33:00.540782", "failed":
true, "rc": 42,
"start": "2018-10-15 11:33:00.536677",
"stderr": "",
"stderr_lines": [], "stdout": "",
"stdout_lines": []}
</pre></td><td>These error messages can be ignored.</td></tr>
</table>

<table>
<tr><th>Issue</th><th>Description</th><th>Resolution or workaround</th></tr>
<tr><td></td><td><pre>2018-10-15 11:33:00,558 p=12895 u=root
| ...ignoring</pre></td><td></td></tr>
</table>

Configure data source connections

Data source connections describe the source data systems for which IBM Spectrum Discover indexes metadata.

Creating data source connections in IBM Spectrum Discover identifies source storage systems that are to be indexed by IBM Spectrum Discover.

For some data source types, a network connection is (optionally) created to allow for automated scanning and indexing of the source system metadata. IBM Spectrum Discover will not index data from unknown sources, so creating a data source connection is the first step towards cataloging any source storage system.

You can add data connections from the source storage systems from the IBM Spectrum Discover graphical user interface.

IBM Spectrum Discover discards any data that comes in from an unknown connection. Therefore, connections must be established before data ingestion. To see the list of defined connections, use the **Data Connection** tab under the **Admin** window of the GUI.

Remember: [If you use a MAC, you might have to adjust the scroll bar settings in **System Settings** to see all available connection types. For example, activate the **Show scroll bars: Always** option.]

Typically, a data source is equivalent to a single file system or object vault or bucket. A data source connection is an alias for the combination of a cluster name and a data source within the cluster. This allows multiple file systems or buckets or vaults with the same name to be indexed by IBM Spectrum Discover when they are in separate clusters.

Note: To create a data source connection, you must have **Data Admin** privileges.

[

Remember: Avoid scanning data concurrently on multiple different data source types.

For example, you can scan multiple similar data source types. Similar data source types include either multiple IBM Spectrum Scale file systems or multiple IBM COS vaults. However, if you scan multiple different data source types concurrently (such as IBM Spectrum Scale file systems and IBM COS vaults together), you might receive:

- System errors
- A potential loss of information during data ingestion

]

IBM Spectrum Scale data source connection

This topic describes how to create an IBM Spectrum Scale source connection, scan a data source, and manually initiate a scan.

Creating an IBM Spectrum Scale data source connection

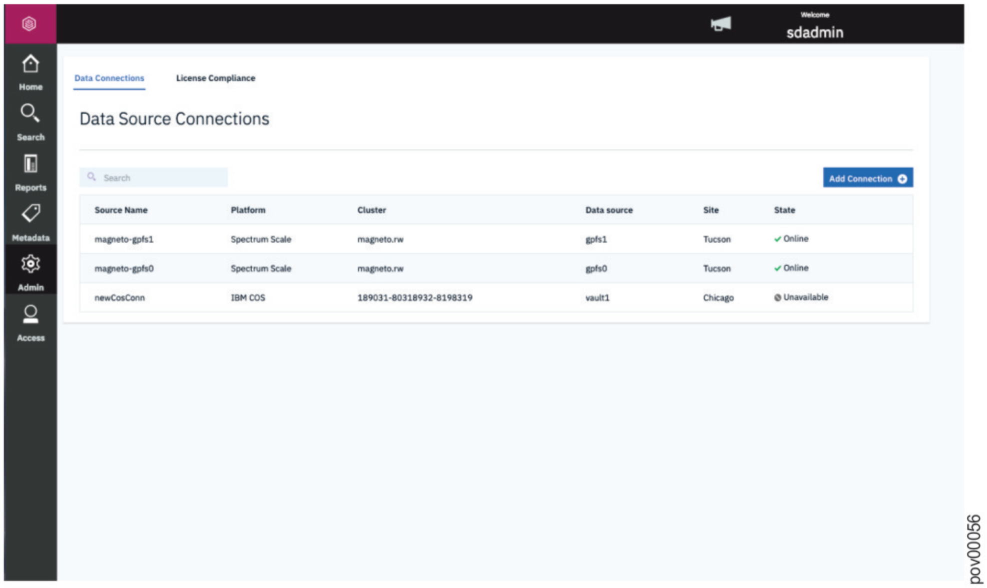
Creating data connections from the source storage systems from the IBM Spectrum Discover graphical user interface.

Procedure

1. Log in to the IBM Spectrum Discover web interface with a user id that has the data admin role associated with it.

   The data admin access role is required for creating connections. For more information about role based access control, see [Managing user access].
2. Select **Admin** from the left navigation menu.

   Click **Admin** to display the different types of data source connection names, platforms, clusters, data source, size, and **Add Connection**.

   Data Source Connections

Source Name	Platform	Cluster	Data source	Site	State
magneto-gpfs1	Spectrum Scale	magneto.rw	gpfs1	Tucson	✓ Online
magneto-gpfs0	Spectrum Scale	magneto.rw	gpfs0	Tucson	✓ Online
newCosConn	IBM COS	189031-80318932-8198319	vault1	Chicago	⊘ Unavailable

   Figure 6. Displaying the source names for data source connections
3. Click **Add Connection** to display a new window that shows **Data Connections Add Data Source Connection**.

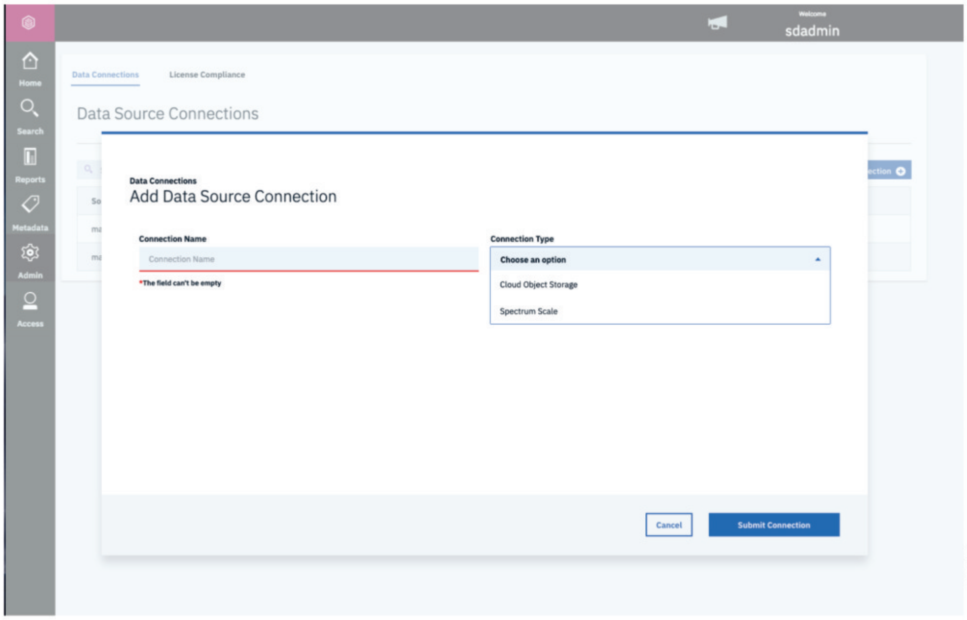


Figure 7. Example of window that shows Data Connections Add data source Connection

4. Do the following steps:
   a) In the field for **Connection Name**, define a **Connection Name**.
   b) Click on the **Connection Type** drop-down menu and **Choose an option** to display the connection type options.
5. Select the connection type to IBM Spectrum Scale.

   Figure 8 on page 81 shows an example of the IBM Spectrum Scale connection.

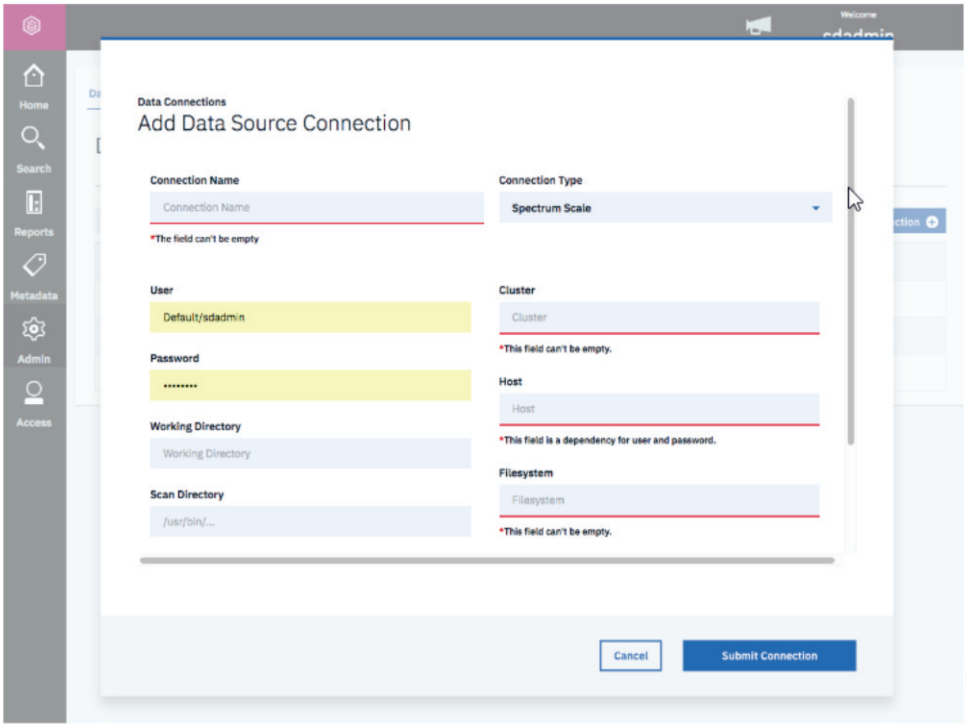


Figure 8. Example of a screen for an IBM Spectrum Scale connection

6. In the screen for IBM Spectrum Scale, fill in the fields, and click **Submit Connection**.

   For IBM Spectrum Scale connections

Connection name

The name of the connection, an identifier for the user, for example `filesystem1`.

Note: It must be a unique name within IBM Spectrum Discover.

User

A user id that has permissions to connect to the data source system and initiate a scan. Go to the following link for the Best Practices Guide for setting up a scan user “Prerequisites for scanning IBM Spectrum Scale systems” on page 86.

Password

The password for the user id specified in user.

Working Directory

A scratch directory on the source data system where IBM Spectrum Discover can put its temporary files.

Scan Directory

The root directory of the scan. All files and directories under this one will be scanned. Typically, this is the base directory of the filesystem, for example `/gpfs/fs1`.

Connection Type

The type of source storage system this connection represents.

Site

An optional physical location tag that an administrator can provide if they want to see the physical distribution of their data.

Cluster

The Scale/GPFS cluster name. To obtain, run the following from the IBM Spectrum Scale file system: `/usr/lpp/mmfs/bin/mmlscluster`.

Host

The hostname or IP address of an IBM Spectrum Scale node from which a scan can be initiated, for example a quorum-manager node.

Filesystem

The short name (omit `/dev/`) of the filesystem to be scanned. For example, `fs1`.

Note: It is important to exactly match the file system name (data source) that IBM Spectrum Scale populates in the scan file. To determine this, run the following command on the IBM Spectrum Scale system: `/usr/lpp/mmfs/bin/mmlsmount all`

Node list

The list of nodes or node classes that will participate in the scan of an IBM Spectrum Scale file system.

Note: When creating data source connections for IBM Spectrum Scale file systems, it is important to exactly match the cluster name and the file system name (data source) that IBM Spectrum Scale populates in the scan file. To determine this, run the following commands on the IBM Spectrum Scale system:

```
/usr/lpp/mmfs/bin/mmlscluster
/usr/lpp/mmfs/bin/mmlsmount all
```

For example:

```
$ /usr/lpp/mmfs/bin/mmlscluster

GPFS cluster information
========================
 GPFS cluster name:         modevvm19.tuc.example.com,
 GPFS cluster id:           7146749509622277333
 GPFS UID domain:           modevvm19.tuc.example.com
 Remote shell command:      /usr/bin/ssh
 Remote file copy command:  /usr/bin/scp
 Repository type:           CCR
Node  Daemon node name           IP address    Admin node name            Designation
----------------------------------------------------------------------------------------
1     modevvm19.tuc.example.com  203.0.113.24  modevvm19.tuc.example.com  quorum-manager
```

```
$ /usr/lpp/mmfs/bin/mmlsmount all
File system gpfs0 is mounted on 1 nodes.
```

IBM Spectrum Scale connection creation readiness list

Use the IBM Spectrum Scale connection creation checklist.

Make sure that you:

- - Identify a node in the target IBM Spectrum Scale cluster to use for the IBM Spectrum Discover connection to the GPFS cluster
  - Identify the node list or node class that participates in the scanning activity
  - Create or identify a user ID and password for scanning
  - Add the scan user to `sudoers` list with `NOPASSD` access for `/usr/lpp/mmfs/bin/mmapplypolicy` and `/usr/lpp/mmfs/bin/mmrepquota`
  - Can use Secure Shell (SSH) to log into the IBM Spectrum Scale system with the scanning user ID and password
  - Run the `sudo /usr/lpp/mmfs/bin/mmapplypolicy` command on the IBM Spectrum Scale node.
  - Validate that the scanning-related working directory:
    - Exists
    - Is globally accessible by the scan worker nodes
    - Scan user has write permissions to the directory (ownership of the directory is preferred but not mandatory)
  - Place the `id_rsa` and `id_rsa.pub` files in `/gpfs/gpfs0/connections/scale/` directory on the IBM Spectrum Discover instance if a specific RSA key pair for passwordless SSH is wanted
  - Validate that these things are installed on the IBM Spectrum Scale node that is identified when you identify a node in the target IBM Spectrum Scale cluster to use for the IBM Spectrum Discover connection to the GPFS cluster:
    - A `python2` level of at least Python 2.7.5
    - A sufficient level of `librdkafka`
    - An appropriate level of `confluent-kafka`

    Note: This is recommended for optimized scan ingestion but is optional.

Automated scan prerequisites

Use IBM Spectrum Scale automated scanning features.

There are two levels of automated scanning of IBM Spectrum Scale systems that IBM Spectrum Scale supports. Both levels require that IBM Spectrum Discovermust establish a passwordless Secure Shell (SSH) connection to the IBM Spectrum Scale clustered system that is being scanned. The difference involves whether the output of the IBM Spectrum Scale policy that is run to do the scan is stored in a file (which then must be automatically copied back to IBM Spectrum Discover and ingested locally) or if the output is instead pushed to the ingest Kafka queue of IBM Spectrum Discover system directly from the IBM Spectrum Scale policy output.

The Kafka queue of IBM Spectrum Discover system is more space-efficient and time-efficient but does have certain dependencies on the IBM Spectrum Scale clustered system that must be met in order to function. TheIBM Spectrum Discover automated scan code determines whether the dependencies are met on the IBM Spectrum Scale clustered system. If the dependencies are met on the IBM Spectrum Scale clustered system, it attempts to scan the system by using the optimized path. If the dependencies are not met on the IBM Spectrum Scale clustered system, it defaults to the file copy path.

The dependencies that must be satisfied on the IBM Spectrum Scale system in order to optimize automated scan ingest from IBM Spectrum Discover are:

- A librdkafka library version 0.11.4 or later
- A Python version 2.7.5 or later
- A confluent-kafka version that is greater than or equal to the installed librdkafka version

An administrator can determine whether librdkafka is installed on the IBM Spectrum Scale node by running the `find /usr -name "*librdkafka*"` or `ls /lib64/librdkafka*` commands. The librdkafka package ships with newer levels of IBM Spectrum Scale on **x86** and **ppc64le** platforms. However, it can also be built from the source code on older levels of IBM Spectrum Scale or **ppc64** platforms. If the IBM Spectrum Scale system runs on Red Hat Enterprise Linux (RHEL) and is connected to a Red Hat Satellite, you can install it by running `yum install librdkafka` as root. You can find source packages of librdkafka here: https://github.com/edenhill/librdkafka

Python 2.7.5 is a commonly installed level on most IBM Spectrum Scale systems. To determine the Python version on a node, run this command: `python --version`. You can use any level of Python that is greater than version 2.7.5. However, Python 3.0.x versions do not work. If no Python 2.7.x level is installed and but you want optimized scanning, use Python 2.7.16 (which is the newest Python 2.7.x level that can be use). Seehttps://www.python.org/downloads/release/python-2716/ for more information.

After you install a sufficient version of Python, you can install confluent-kafka by using pip. To get pip, you must install the python-setuptools package, which provides a binary called easy_install. See https://pypi.org/project/setuptools/#files for more information on python-setuptools. After easy_install is available, you can install pip by running `easy_install-2.7` pip as root. After you install pip, you can install confluent-kafkaby running `pip install confluent-kafka` as root.

Scanning an IBM Spectrum Scale data source

As an administrator, you can initiate an IBM Spectrum Scale scan from an IBM Spectrum Scale to collect system metadata from an IBM Spectrum Scale file system

About this task

When a scan is initiated from the IBM Spectrum Discover graphical user interface, the data moves asynchronously back to the IBM Spectrum Discover.

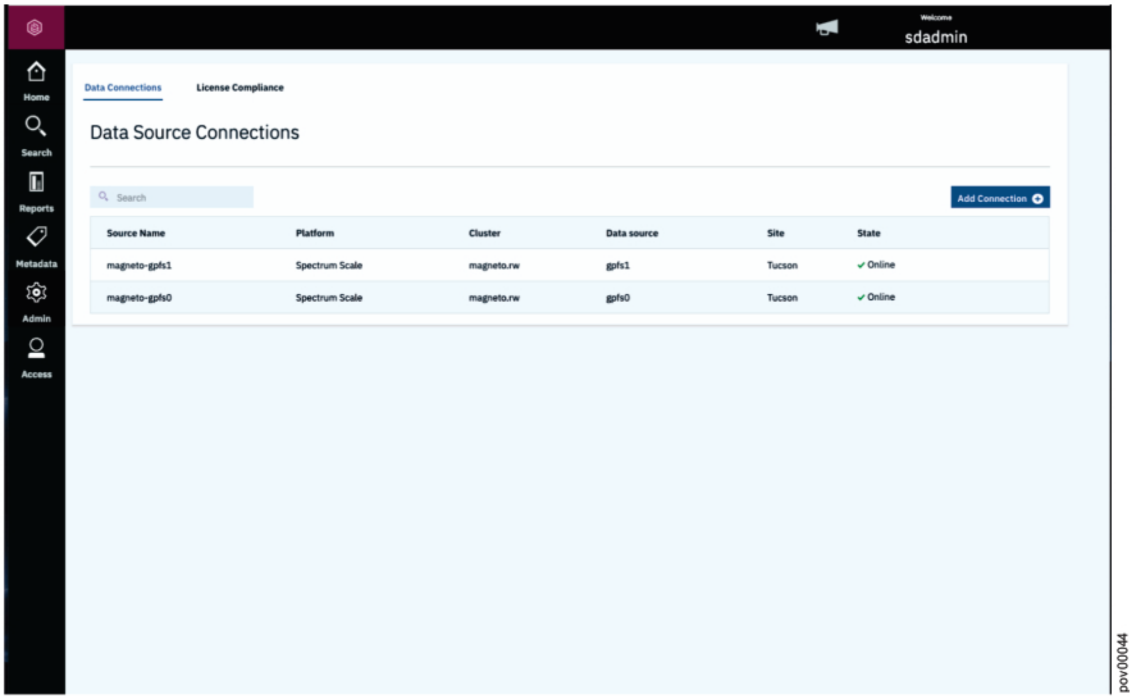
Remember: Before initiating a scan, see "Prerequisites for scanning IBM Spectrum Scale systems" on page 86.

Automated scanning and data ingestion relies on an established and active network connection between the IBM Spectrum Discover instance and the source IBM Spectrum Scale management node. If the connection cannot be established, the state of the data source connection will show as 'unavailable' and the option for automated scanning will not appear in the IBM Spectrum Discover GUI for that connection.

Procedure

1. Go to the IBM Spectrum Discover graphical user interface.
2. Under **Admin**, select **Data Source Connections**.

   Figure 9 on page 85 shows an example of the Admin data connections menu page.



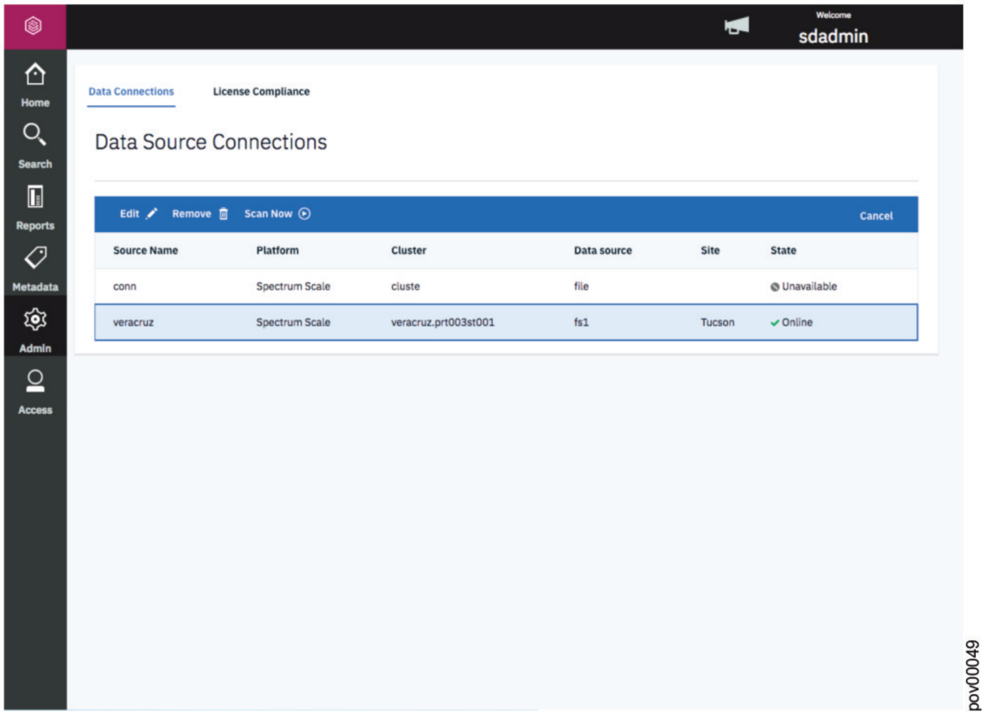
Data Source Connections

Source Name	Platform	Cluster	Data source	Site	State
magneto-gpfs1	Spectrum Scale	magneto.rw	gpfs1	Tucson	✓ Online
magneto-gpfs0	Spectrum Scale	magneto.rw	gpfs0	Tucson	✓ Online

Figure 9. Admin data connections menu page

3. Select the data source connection you want to scan. Ensure that the **State** is listed as **Online** to make your system scan ready.

   Figure 10 on page 85 shows an example of how to connect to the IBM Spectrum Discover library.



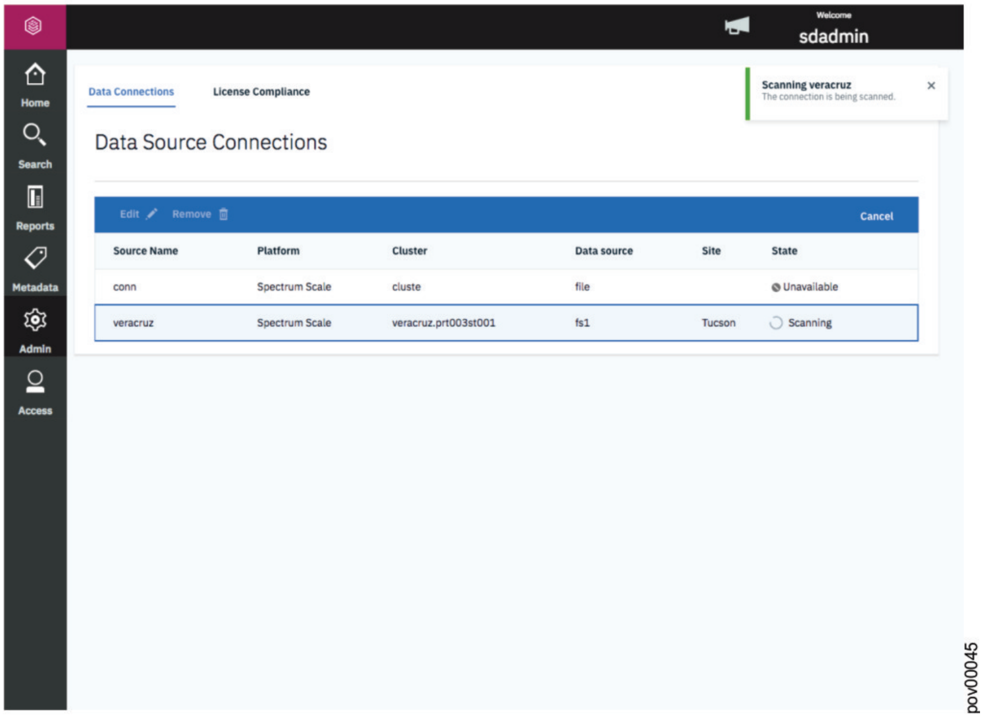
Data Source Connections

Source Name	Platform	Cluster	Data source	Site	State
conn	Spectrum Scale	cluste	file		Unavailable
veracruz	Spectrum Scale	veracruz.prt003st001	fs1	Tucson	✓ Online

Figure 10. How to connect to the IBM Spectrum Discover library

4. Select **Scan Now** to change the status to **Scanning**.

   Figure 11 on page 86 shows an example of a scan that is in a state of **Scanning**.



Scanning veracruz
The connection is being scanned.

Data Source Connections

Source Name	Platform	Cluster	Data source	Site	State
conn	Spectrum Scale	cluste	file		Unavailable
veracruz	Spectrum Scale	veracruz.prt003st001	fs1	Tucson	Scanning

Figure 11. A scan in a state of Scanning

When the scan finishes, the state field returns to a status of **Online**.

Remember: You can also specify a time to begin the scan. Any time zones specified default to Coordinated Universal Time (UTC) time. So, if you specify your scan for 12:00pm, it is 12:00pm UTC.

Prerequisites for scanning IBM Spectrum Scale systems
There are prerequisites for scanning an IBM Spectrum Scale instance with IBM Spectrum Discover in a secure and performant way.

Creating a user ID for scanning
Use this information to create a user ID to scan a system connection.

About this task

Follow these steps on the IBM Spectrum Scale system to create a special user ID for scanning.

Procedure

1. Login to a IBM Spectrum Scale management node as **`root`**.

   Alternatively, you can **sudo** to `root` from another user ID.
2. Use the following `adduser` steps to ensure that you are able to ssh into the cluster:
   a) `adduser` <user> `-m`
   b) `passwd` <user>
3. Run: **`visudo`**
   a) Add this line in the users section: <user> `ALL=NOPASSWD: /usr/lpp/mmfs/bin/mmapplypolicy, /usr/lpp/mmfs/bin/mmrepquota, /usr/bin/easy_install, /usr/bin/pip`
   b) Write and quit: `:wq`
4. Create a IBM Spectrum Discover working directory and ensure that <user> has write permissions.

   For example: `mkdir -p /gpfs/fs1/sd_scan -m 770; chown <user> /gpfs/fs1/sd_scan`

Security considerations
Use this information to securely scan a system connection.

Scanning an IBM Spectrum Scale instance involves utilizing the **mmapplypolicy** command on the IBM Spectrum Scale system, which requires superuser permissions. When creating the data source connection for the target IBM Spectrum Scale system in the IBM Spectrum Discover interface, you are prompted for a *userid* and *password* to enable automated scans. You are not required to provide these credentials if scans will only be run manually on the target IBM Spectrum Scale system by an administrator. However, if automation and/or scheduling of scans is desired, then the authentication credentials are required. When login credentials are provided, IBM Spectrum Discover will attempt to establish a shared-RSA key relationship with the IBM Spectrum Scale system to allow for password-less ssh/sftp between the two systems. The default key pair is generated during deployment and is unique per IBM Spectrum Discover instance. However, you can supply your own key pair (private/public) if desired by overwriting the **id_rsa** and **id_rsa.pub** files in `/gpfs/gpfs0/connections/scale` prior to creating data source connections on IBM Spectrum Discover.

Note: Currently, only a single key pair is used for all the IBM Spectrum Scale connections. If the key files are replaced after a connection has already been created, the existing connection will lose the ability to run automated scans until it is deleted/re-created with the new key pair files in place.

Rather than providing root login credentials, an administrator should create a special user ID with limited permissions on the IBM Spectrum Scale system and enable a password-less **sudo** for the user ID, to the binaries needed for scanning. This will prevent someone from gaining root access to the target IBM Spectrum Scale system if the IBM Spectrum Discover system is somehow compromised.

Follow these steps on the system to create a special user ID for scanning:

1. Login to a management node as `root` (or su to `root` from another user ID).
2. Run the **adduser <user> -m** command. For example:

   ```
   adduser sdadmin -m
   ```

3. Set the Secure Shell (SSH) password for the new user by running the **passwd <user>** command. For example:

   ```
   passwd sdadmin
   ```

4. Run `visudo`.

   a. Add this line in the section for the new user:

   ```
   <user from step 2> ALL=NOPASSWD: /usr/lpp/mmfs/bin/mmapplypolicy, /usr/lpp/mmfs/bin/
   mmrepquota
   ```

   2. Use the **:wq** command to write and quit.
5. Create an SD working directory and make sure that `<user>` has write permissions. For example:

   ```
   mkdir -p /gpfs/fs1/sd_scan -m 770; chown <user> /gpfs/fs1/sd_scan
   ```

Changing passwordless SSH keys

You can rotate RSA authentication key pairs for passwordless SSH on a frequency and remove old security keys from the `authorized_hosts` file on the IBM Spectrum Scale node that IBM Spectrum Discover connects to. To update the authentication keys, follow these steps:

1. Make sure that the `id_rsa.pub` contents for the new authentication key pair are in the `~/.ssh/authorized_hosts` file for the user that is specified in the IBM Spectrum Discover connection document for the IBM Spectrum Scale target file system.
2. Replace the `/gpfs/gpfs0/connections/scale/id_rsa` and `id_rsa.pub` files with the new authentication key pair.

After you replace the authentication key pair, IBM Spectrum Discover uses the authentication key pair to connect to the IBM Spectrum Scale target system.

Performance considerations

Use this information to scan a system connection without degrading performance.

Running a scan policy on an IBM Spectrum Scale system can be resource intensive and cause noticeable performance degradation on the IBM Spectrum Scale system. Often, system administrators choose to designate certain nodes or node classes for running the scans. The IBM Spectrum Discover interface has an input field when creating IBM Spectrum Scale connections for the administrator to specify which nodes or node class(es) they would like to run the scan on. The value `all` will run the scan across all nodes in the cluster. Any other list (comma separated) will be treated as a list of nodes or node classes on which to run the scan. Scan times vary by the size of the filesystem, how many nodes are used in the scan, how many CPUs are used per node, and whether or not the IBM Spectrum Scale cluster metadata is in flash memory.

Manually initiating an IBM Spectrum Scale scan

How to configure IBM Spectrum Discover to connect to IBM Spectrum Scale. After completing these steps, data can be ingested from an IBM Spectrum Scale data source to IBM Spectrum Discover for metadata indexing.

Before you begin

Create the data source connection to IBM Spectrum Scale. For more information, see “Configure data source connections” on page 79.

Restriction: IBM Spectrum Discover uses a unit separator (ASCII code 0x1F) as the field delimiter for ingestion into the database. This means that data which contains this character in path/file/object names results in improper parsing of the input data and the records are rejected by IBM Spectrum Discover.

Procedure

1. Perform a file system scan to collect system metadata from IBM Spectrum Scale to be ingested into IBM Spectrum Discover. For more information, see “Performing file system scan to collect metadata from IBM Spectrum Scale” on page 88.
2. Copy the output of the file system scan to the IBM Spectrum Discover master node. For more information, see “Copying the output of the IBM Spectrum Scale file system scan to the IBM Spectrum Discover master node” on page 91.
3. Ingest data from the file system scan in IBM Spectrum Discover. For more information, see “Ingesting metadata from IBM Spectrum Scale file system scan in IBM Spectrum Discover” on page 92.
4. Ingest quota information from the file system. For more information, see “Ingesting quota information from the file system” on page 92.

Performing file system scan to collect metadata from IBM Spectrum Scale

You can use the file system scanning tool, IBM Spectrum Scale Scanner, to collect system metadata from IBM Spectrum Scale to be ingested into IBM Spectrum Discover.

About this task

The IBM Spectrum Scale Scanner tool uses the IBM Spectrum Scale information lifecycle management (ILM) policy engine to obtain the system metadata about the files stored on the file system. The system metadata is written to a file and the file is transferred to the IBM Spectrum Discover master node where it is ingested and analytics is performed to provide search, duplicate file detection, archive data detection, and capacity show-back report generation. The following system metadata is collected from the file system scan:

Key name	Description
site	The site where the file or object resides
platform	The source storage platform that contains the file or object
size	The size of the file
owner	The owner of the file
path	The sub-directory where the data resides
name	The name of the data
permissions	The permissions for the file (mode)
ctime	The change time of the file metadata (inode)
mtime	The time when the data was last modified
atime	The time when the data was last accessed
Filesystem	The name of the IBM Spectrum Scale file system that is storing the data
Cluster	The name of the IBM Spectrum Scale cluster
inode	The IBM Spectrum Scale inode that is storing the data
Group	The Linux group associated with the file
Fileset	The fileset that is storing the file
Pool	The storage pool where the file resides
Migstatus	If applicable, indicates whether or not the data is migrated to tape or object
migloc	If applicable, indicates the location of the data if migrated to tape or object
ScanGen	Scan generation - useful to track re-scans

The IBM Spectrum Scale Scanner tool also collects quota information by calling **mmrepquota**.

The tool comprises the following files:

- scale_scanner.py: The tool that invokes the IBM Spectrum Scale ILM policy
- scale_scanner.conf: The configuration file used to customize the behavior of the scale_scanner.py tool
- createScanPolicy: The script that is called internally by the tool

Procedure

Install the IBM Spectrum Scale Scanner tool by unpacking the utility from the IBM Spectrum Discover node to the desired location on the IBM Spectrum Scale cluster node.

1. Login to the IBM Spectrum Discover node through Secure Shell (SSH) with the moadmin username and password:

```
ssh modadmin@spectrum.discover.ibm.com
```

2. Change to the directory that contains the Spectrum Scale scanning utility:

```
/opt/ibm/metaocean/spectrum-scale/etc/metaocean
```

3. scp the `createScanPolicy`, `_init_.py`, `scale_scanner.conf`, and `scale_scanner.py` files to a node in the IBM Spectrum Scale cluster:

```
scp * root@spectrumscale.ibm.com:/my_scanner_directory
```

```
createScanPolicy 100% 3320 3.2KB/s 00:00
init.py 100% 427 0.4KB/s 00:00
scale_scanner.conf 100% 1595 1.6KB/s 00:00
scale_scanner.py 100% 13KB 13.2KB/s 00:00
```

4. On the IBM Spectrum Scale node where you install the scanning utility, edit the configuration file (`scale_scanner.conf`) as follows:

   a) Use the IBM Spectrum Discover UI to create a connection to the SS system on which you start a manual scan for. Set the `filesystem` and `scandir` fields, and optionally set the `outputdir` and `site` fields in the `[spectrumscale]` stanza of the file.

```
[spectrumscale]
# Spectrum Scale Filesystem which hosts the scan directory
# example:  /dev/gpfs0
filesystem=/dev/gpfs0
# The directory path on Spectrum Scale Filesystem to perform scan on
# example: /gpfs0
# specifies a global directory to be used for temporary storage during
# mmapplypolicy command processing. The specified directory must be
#mounted with read/write access within a shared file system
mountpoint=mount point of the gpfs filesystem
# It is unclear what the mount_point should be, but setting the mount point
# to the mount point of the scale file system on the IBM Spectrum Scale node works.
scandir=/gpfs0
# The directory to store output data from the scan in (default is
# scandir)
outputdir=
# The site tag to specify a physical location or organization identifier.
# If you use this field, remove the comment (#)
#site=
```

   b) Set the `scale_connection`, `master_node_ip`, and `username` fields in the `[spectrumdiscover]` stanza of the file.

   Note: `scale_ connection` refers to the name of the IBM Spectrum Scale file system that will be scanned and ingested into IBM Spectrum Discover. The `scale_connection` value must match the value defined in the `Data  Source` column of the **Data Connections** page in the IBM Spectrum Discover GUI.

   The username must be a valid user name in IBM Spectrum Discover that has the `dataadmin` role. The **username** field takes the format of `<domain_name>/<username>`.To determine a domain and username with the `dataadmin` role, go to the **Access Users** page in the IBM Spectrum Discover GUI and click on view for the defined users.

   In the case of the local domain, it is not necessary to specify the domain as part of the username field as this is the default domain. For example, given a user name of user1 in the local domain that has been assigned the `dataadmin` role, in the configuration file enter the following value: `username=user1`

```
[spectrumdiscover]
# Name of the Spectrum Scale connection to scan files from
# Check using the Spectrum Discover connection manager APIs
scale_connection=fs3
# Spectrum Discover Master Node IP
master_node_ip=203.0.113.23
# Spectrum Discover user name, having 'dataadmin' role
# Use format <domain_name>/<username>
# e.g. username=Scale/scaleuser1
username=user1
```

Note: The scanner output file generates approximately 1K of metadata for every file in the system. If there are 12M files, the size is expected to be approximately 12GB. By default, the output file is written to the same directory that is being scanned. The log file output location can be customized by setting the `outputdir` field.

5. Run the scan by using the following command:

```
./scale_scanner.py
```

Note: While running the **`./scale_scanner.py`** command, you can start another scan. If you start another scan, ensure that you run the scan with another connection that is online and is not being scanned currently. When the scanner is running, the scanner hides the **scan now** button automatically.

Note: As you run the `scale_scanner.py` script, you are prompted for the password for the IBM Spectrum Discover user that you have configured in the `scale_scanner.conf` file with the `username` under the `spectrumdiscover` section. You must provide the correct password for the configured user. As described in the configuration file, this user needs to be a valid user configured in the IBM Spectrum Discover Authentication service (Access management). Also, this user must have the `dataadmin` role assigned.

For example:

```
$ ./scale_scanner.py
Enter password for SD user 'user1':
Scale Scan Policy is created at: ./scanScale.policy
```

Note:

- After you see a line similar to “0 ‘skipped’ files and/or errors” press enter to return to the command prompt.
- The scan takes about 2 minutes 30 seconds for every 10M files on the following configuration:

```
x86 -based Spectrum Scale Cluster
•4 M4 NSD client nodes
•2 M4 NSD server nodes
•DCS3700 350 2TB NL SAS drives & 20 200GB SSD
•QDR InfiniBand cluster network
```

Copying the output of the IBM Spectrum Scale file system scan to the IBM Spectrum Discover master node

After you have scanned your IBM Spectrum Scale file system and have the `list.metaOcean` output file, copy it to the IBM Spectrum Discover master node.

Procedure

As an IBM Spectrum Discover administrator, use **scp** to copy `list.metaOcean` file from the scan output directory to the `/gpfs/gpfs0/producer` directory on the master node.

Note: If there are multiple file systems in the same cluster that are being scanned, you can rename the `list.metaOcean` file to avoid name conflicts and to not overwrite an existing `list.metaOcean` file that is in use. For example:

```
$ mv list.metaOcean list.metaOcean.myfilesystem
$ scp list.metaOcean.myfilesystem moadmin@MasterNodeIP:/gpfs/gpfs0/producer
```

Ingesting metadata from IBM Spectrum Scale file system scan in IBM Spectrum Discover
Records are inserted into IBM Spectrum Discover for indexing when they are pushed to a Kafka connector topic corresponding to the type of data being ingested. In the case of IBM Spectrum Scale, the Kafka connector topic type is `scale-scan-connector-topic`.

About this task

A Kafka client producer is required to put the IBM Spectrum Scale file system scan file records onto the Kafka connector topic. The following steps show how to use the **`ingest`** alias command to push the records in the `list.metaOcean` file (or another named file) onto the Kafka connector topic.

Procedure

1. Run the following command to ingest the data:

   ```
   $ ingest /gpfs/gpfs0/producer/list.metaOcean
   ```

2. Replace the `list.metaOcean` path with the path of the file that you want to ingest.

Ingesting quota information from the file system
The file system scanning tool, IBM Spectrum Scale Scanner, has the ability to harvest and send quota information to IBM Spectrum Discover.

Procedure

To perform quota ingestion, run the following command on the IBM Spectrum Scale cluster node:

```
./scale_scanner.py --quota-only
```

For example:

```
$ sudo ./scale_scanner.py --quota-only
Enter password for SD user 'user1':
```

IBM Cloud Object Storage data source connection

You can create a IBM Cloud Object Storage (COS) connection and initiate a scan.

IBM COS uses a connector residing on the storage system to push events to a Kafka topic residing in the IBM Spectrum Discover cluster. When configured, the IBM Spectrum Discover consumes the events and indexes them into the IBM Spectrum Discover database.

Restriction: IBM Spectrum Discover uses a unit separator (ASCII code 0x1F) as the field delimiter for ingestion into the database. This means that data which contains this character in path/file/object names results in improper parsing of the input data and the records are rejected by IBM Spectrum Discover.

Creating an IBM Cloud Object Storage data source connection
You can create an IBM Cloud Object Storage (COS) data source connection and from the storage system.

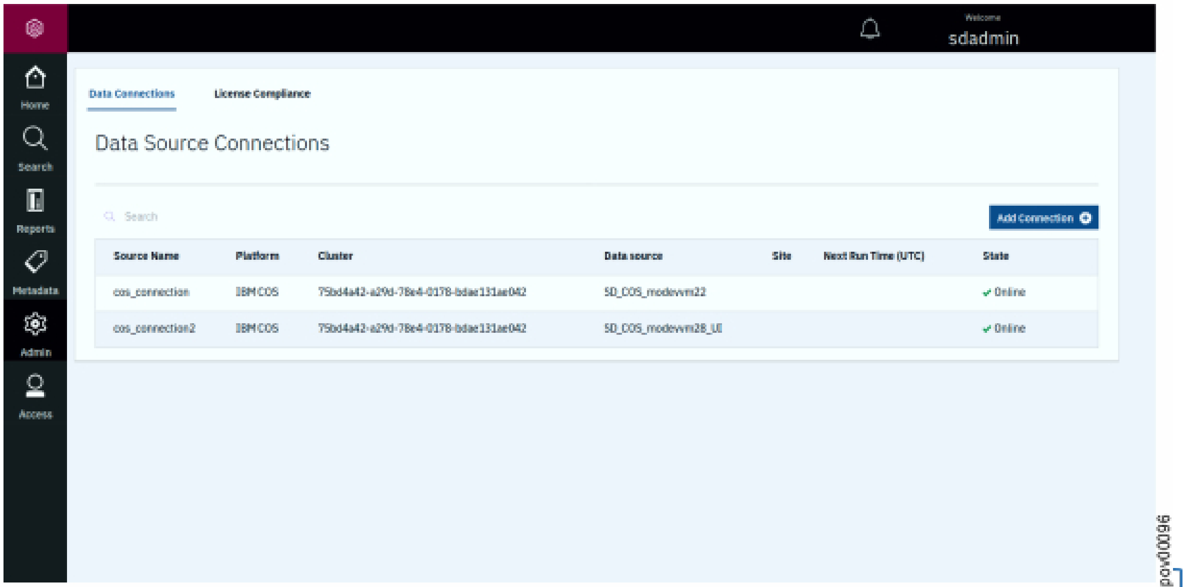
Procedure

1. Log in to the IBM Spectrum Discover web interface with a user ID that has the data admin role associated with it.

   The data admin access role is required for creating connections. For more information about Role-Based Access Control (RBAC), go to https://www.ibm.com/support/knowledgecenter/SSY8AC_2.0.0/isd200_welcome.html.

2. Select **Admin** from the left navigation menu.

   Clicking **Admin** displays the different types of data source connection names, platforms, clusters, data source, size, and **Add Connection**.



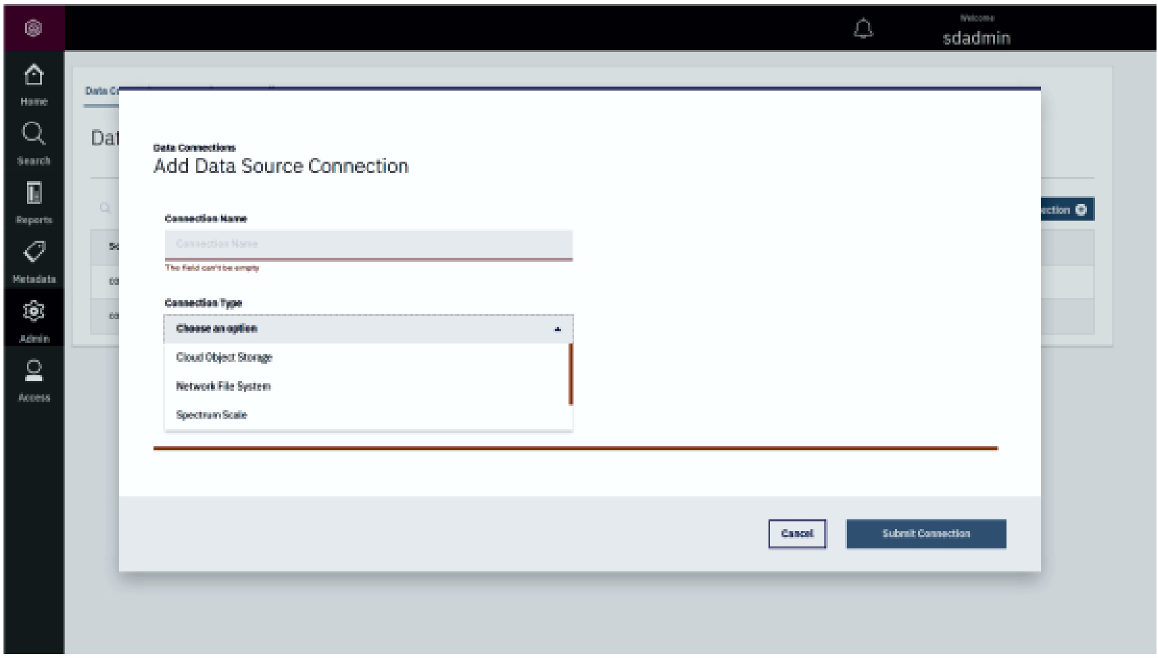
Data Source Connections

Search

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
cos_connection	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevm22			✓ Online
cos_connection2	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevm28_UI			✓ Online

Figure 12. Displaying the source names for data source connections

3. Click **Add Connection** to display a new window that shows **Data Connections Add Data Source Connection**.



Data Connections

Add Data Source Connection

Connection Name

Connection Name

The field can't be empty

Connection Type

Choose an option

Cloud Object Storage

Network File System

Spectrum Scale

Figure 13. Example of window that shows Data Connections Add data source Connection

4. Do the following steps:

   a) In the field for **Connection Name**, define a **Connection Name**.

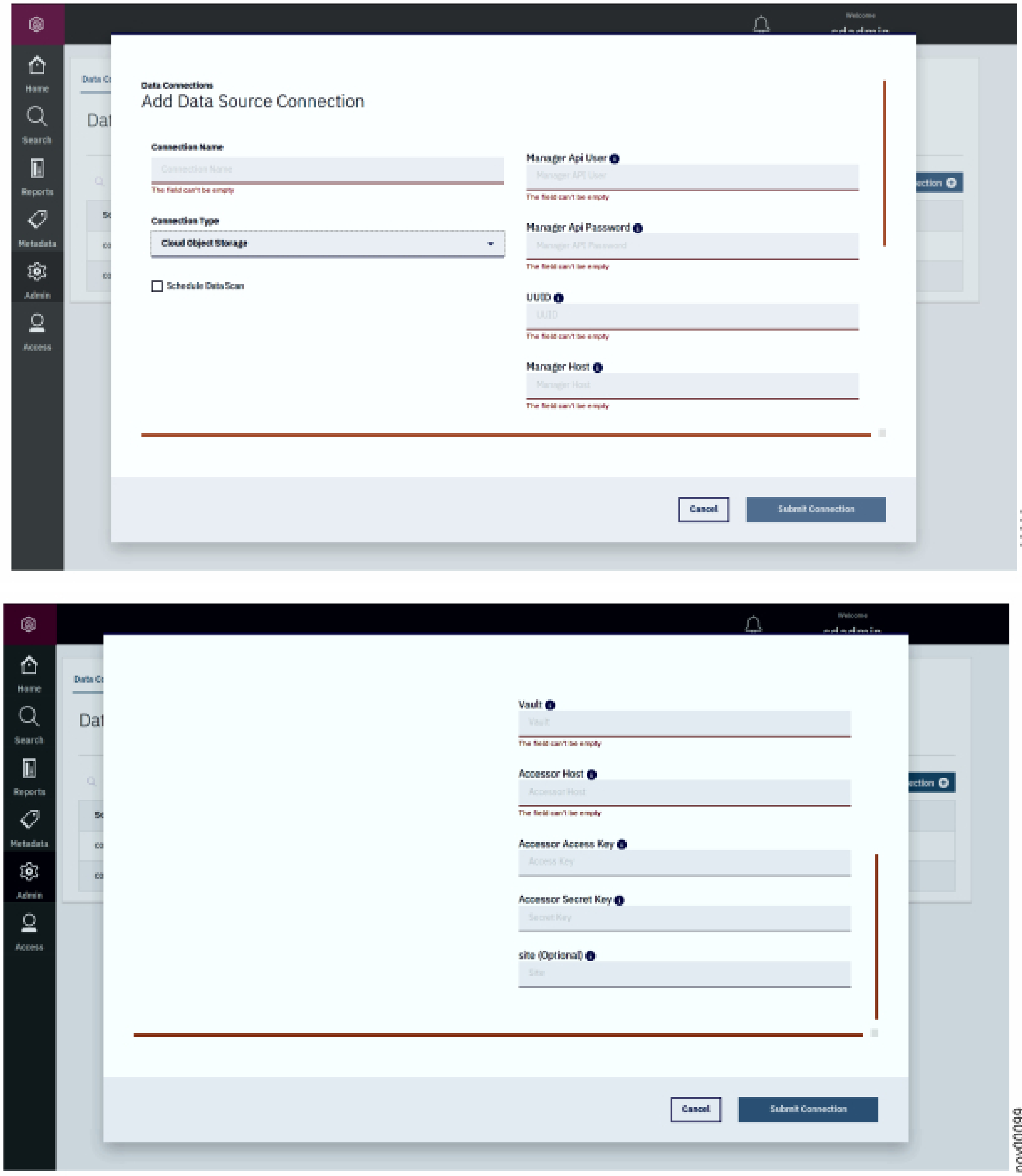
   b) [

   Click the down arrow for **Connection Type** to display a drop-down menu for the connection type.

   ]

5. Select the connection type **Cloud Object Storage**.

   Figure 14 on page 94 shows an example of the screen for an IBM COS connection.



Data Connections

Add Data Source Connection

Connection Name

Connection Name

The field can't be empty

Connection Type

Cloud Object Storage

□ Schedule Data Scan

Manager Api User

Manager API User

The field can't be empty

Manager Api Password

Manager API Password

The field can't be empty

UUID

UUID

The field can't be empty

Manager Host

Manager Host

The field can't be empty

Cancel

Submit Connection

Vault

Vault

The field can't be empty

Accessor Host

Accessor Host

The field can't be empty

Accessor Access Key

Access Key

Accessor Secret Key

Secret Key

site (Optional)

Site

Cancel

Submit Connection

Figure 14. Example of the screen for a IBM COS connection

6. In the screen for **Cloud Object Storage**, complete fields, and click **Submit Connection**.

For Cloud Storage Object connections Manager

Manager API user

A user ID that has permissions to connect to the data source system.

Manager API Password

The password for the user ID specified above.

UUID

The unique id of the DSNet cluster. To obtain the UUID, log in to the COS Manager GUI and click **Help** > **About this system** on the upper-right corner of the window.

Host

The IP or hostname of the manager node within the DSNet.

Vault

The specific data vault represented by this connection.

Site

An optional physical location tag that an administrator can provide if they want to see the physical distribution of their data.

Accesser

The IP address or hostname of the accesser node on DSNet.

Accesser access key

The accesser access key that has permission to access data in the data vault that is to be scanned. If the accessor access key value is blank, the value is retrieved (for the manager API user) from the manager API.

Accesser secret key

The accesser secret key that has permission to access data in the vault that is to be scanned. If the secret access key value is blank, the value is retrieved (for the manager API user) from the manager API.

Scanning an IBM Cloud Object Storage data connection

You can initiate an IBM connection scan to collect system metadata from an IBM Cloud Object Storage (IBM COS) system.

About this task

When you initiate a scan from the IBM Spectrum Discover graphical user interface (GUI), the metadata is transferred asynchronously back to the IBM Spectrum Discover instance.

Note:

IBM Spectrum Discover does not support scanning of vaults in a dsNet that has any of the following:

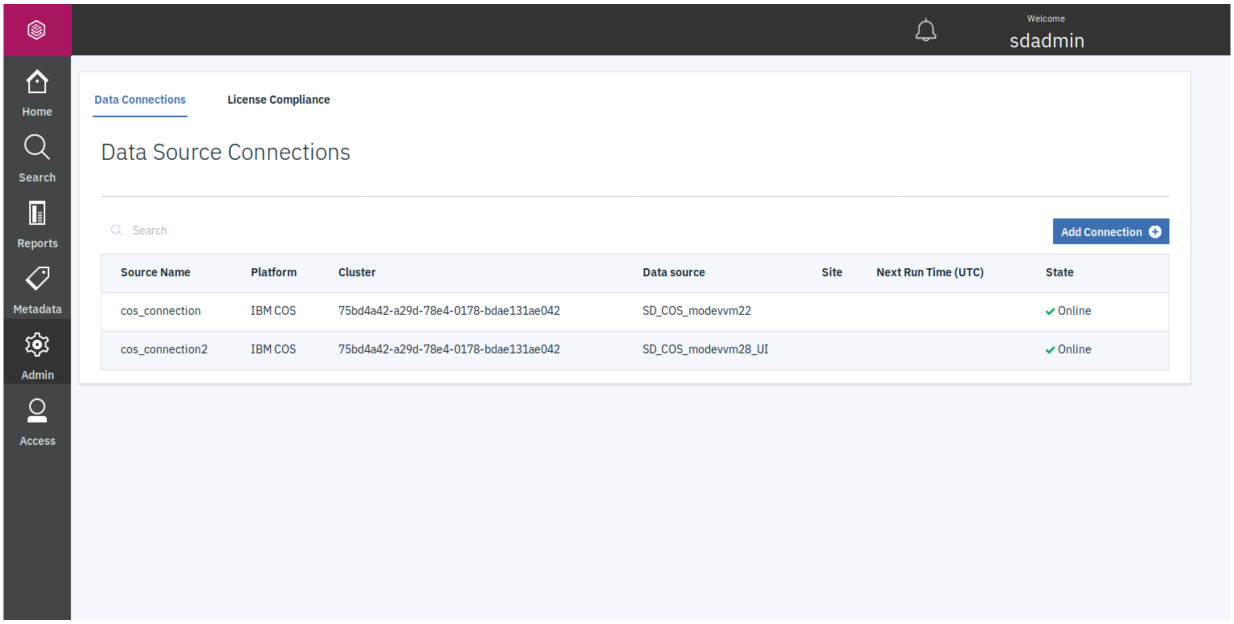
- Proxy vault
- Mirrored vault
- Vault setup for migration

Automated scanning and data ingestion relies on an established and active network connection between the IBM Spectrum Discover instance and the COS storage source. If the connection cannot be established, the state of the datasource connection shows as unavailable, and the option for automated scanning does not appear in the IBM Spectrum Discover GUI for that connection.

Procedure

1. Log into the IBM Spectrum Discover graphical user interface (GUI).
2. Under **Admin** select **Data Source connections**

   The following example shows the **Admin** data connections menu page:



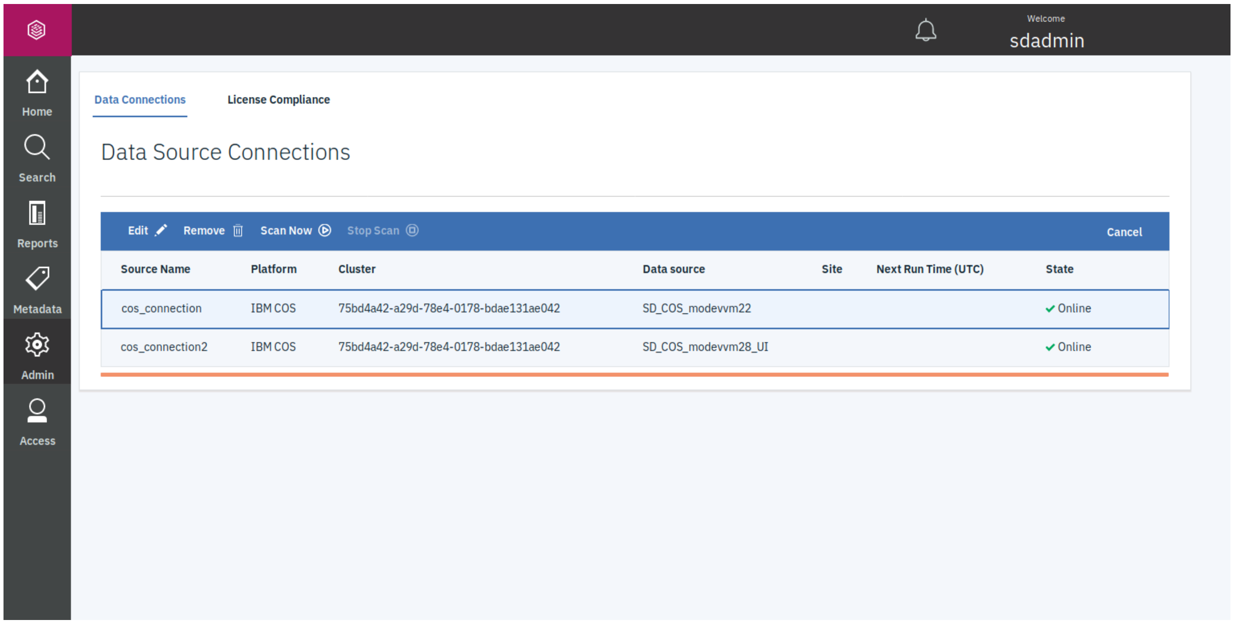
Data Source Connections

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
cos_connection	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevvm22			✓ Online
cos_connection2	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevvm28_UI			✓ Online

Figure 15. Data source connections

3. Select the data source connection you want to scan. Ensure that the `State` is listed as `Online` to make your system scan ready.

   The following example shows how to connect to the IBM Spectrum Discover library.



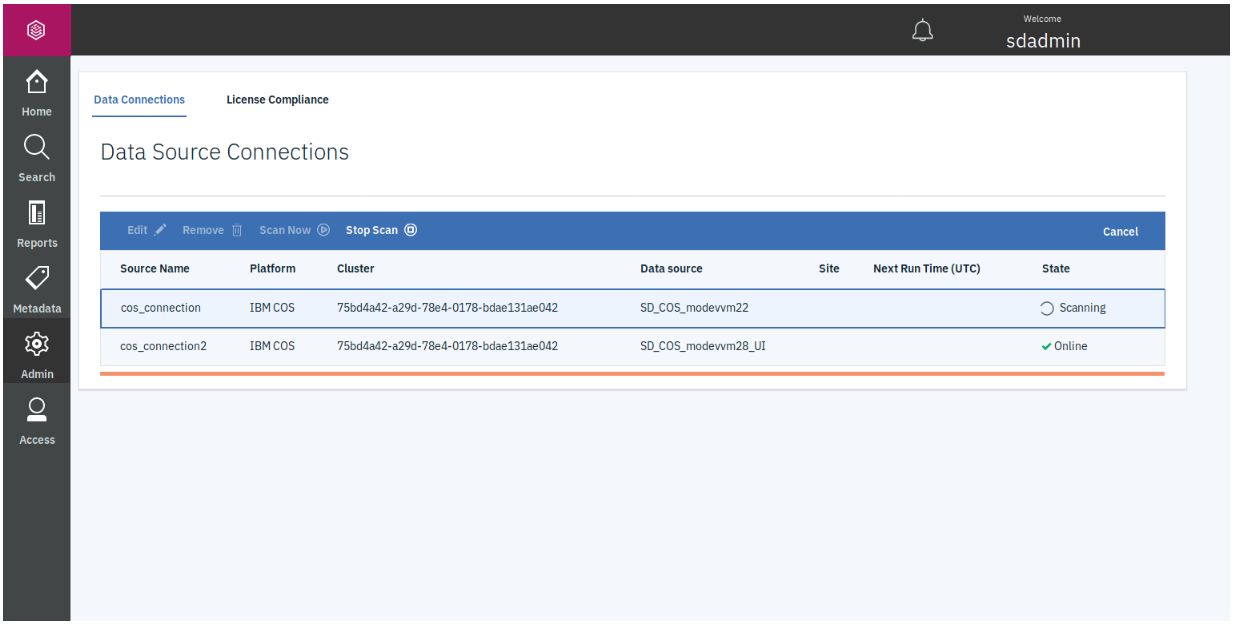
Data Source Connections

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
cos_connection	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevvm22			✓ Online
cos_connection2	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevvm28_UI			✓ Online

Figure 16. Selecting a data source connection to scan

4. Select **Scan Now** to change the status to **Scanning**.

   The following example shows an active scan.



Data Source Connections

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
cos_connection	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevvm22			Scanning
cos_connection2	IBM COS	75bd4a42-a29d-78e4-0178-bdae131ae042	SD_COS_modevvm28_UI			✓ Online

Figure 17. Active scans

5. When the scan finishes, the state field returns to a status of **Online**.

Best practices for scanning IBM Cloud Object Storage systems

Use best practices for scanning IBM Cloud Object Storage (IBM COS) systems .

It is recommended to check the log files in the following directories after each scan:

`/gpfs/gpfs0/connections/cos/<connection_name>/debug/<scan_timestamp>/scanner.debug` indicates whether the scan was successful or not.

`/gpfs/gpfs0/connections/cos/<connection_name>/error/<scan_timestamp>/scanner.error` contains a list of all the messages that are not delivered to IBM Spectrum Discover

This file contains a list of all the messages that are not delivered to IBM Spectrum Discover.

`/gpfs/gpfs0/connections/cos/<connection_name>/data/<scan_timestamp>/` contains a subfolder with the scanned data source name. There is a stats folder inside this folder that contains information about the number of objects in the data source or the number of objects or scanned files.

You can also compare the total size of the bucket that is reported in IBM Spectrum Discover with the total size of the IBM COS at its source (if it is available).

Replaying IBM Cloud Object Storage notifications

Use the IBM Cloud Object Storage (COS) Replay feature to resend notifications that failed because of an outage or loss of data.

The IBM COS Replay reads object metadata from vaults and submits the metadata to IBM Spectrum Discover by using Kafka notifications.

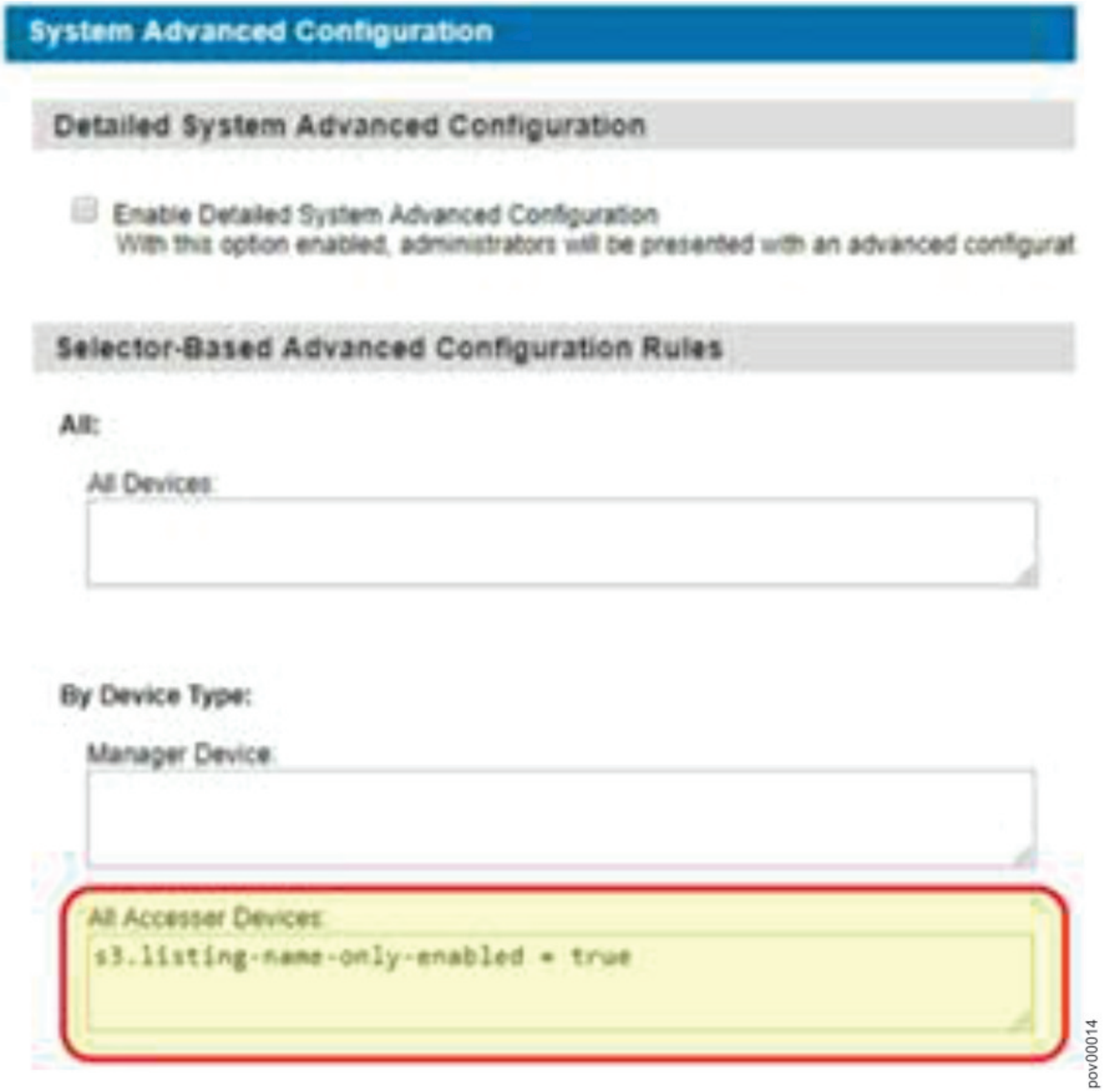
Prerequisites

The IBM Cloud Object Storage Scanner prerequisites are listed in this topic:

- You must enable the Get Bucket Extension for all accessor devices.

  To enable the Get Bucket Extension, you must set the s3.listingname-only-enabled equal to true in the Manager System Advanced Configuration.
- `Access_logs` are uploaded to the management vaults up to 15 minutes after roll-over. Roll-over can be triggered earlier by setting the Rotation Period to 15 minutes in Manager under Maintenance/Logs/Device Log Configuration. Refer to IBM Cloud Object Storage documentation to ensure this is configured and the relevant access logs are present prior to executing the Replay.

See Figure 18 on page 98.



System Advanced Configuration

Detailed System Advanced Configuration

□ Enable Detailed System Advanced Configuration
With this option enabled, administrators will be presented with an advanced configurat

Selector-Based Advanced Configuration Rules

All:

All Devices:

By Device Type:

Manager Device

All Accesser Devices:

```
s3.listing-name-only-enabled = true
```

Figure 18. Example of the system advanced configuration

Note: You do not need to restart the Accessor, but if you do not restart the Accessor, you might need to wait for 5 minutes before the setting takes effect.

Overview of architecture

This topic describes a high-level overview of IBM Cloud Object Storage Scanner architecture.

The following figure shows a high-level overview of IBM Cloud Object Storage Replay architecture.

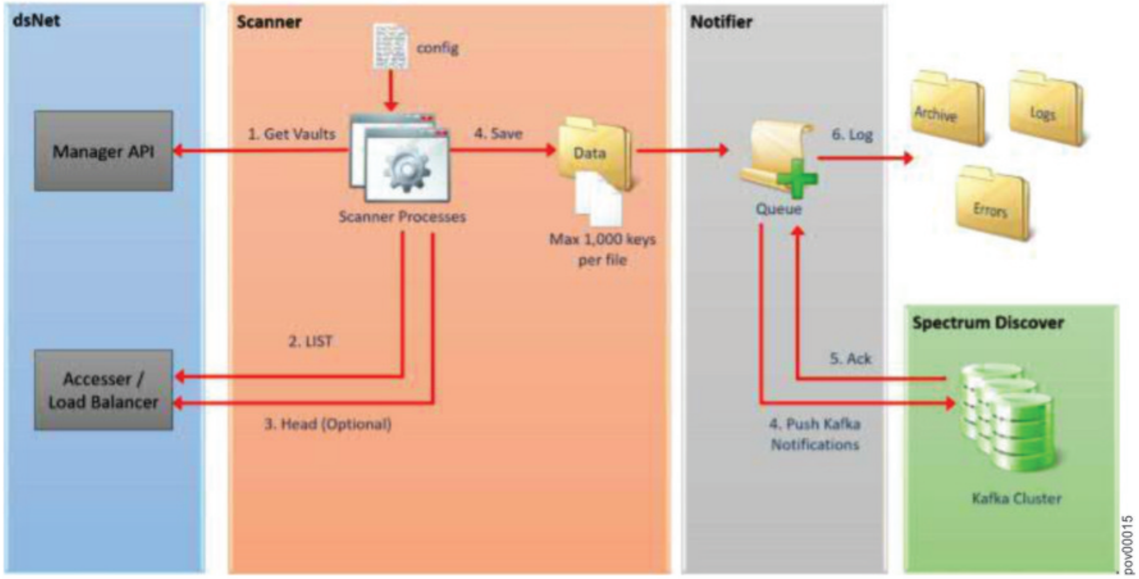


Figure 19. IBM Cloud Object Storage [Replay] replay architecture

[The `/gpfs/gpfs0/connections/cos/replay/output/data` folder puts the replay output and the notifier reads Kafka messages from the directory. Putting replay output onto disk means that a cold restart is possible.]

The IBM Cloud Object Storage [Replay] consists of [two] major components:

Replay
Downloads system's logs and re-creates notifications that are sent during a defined time period.

Notifier
Submits the extracted information to IBM Spectrum Discover.

Configuration file

[The configuration file is used by the Notifier and Replay.]

The configuration file includes:

- Information regarding the net
- [Runtime parameters for the Notifier and Replay]
- A list of vaults to scan

[The configuration file is named `scanner-settings.json` and must sit in the `/gpfs/gpfs0/connections/cos/replay` directory.]

The rules for IBM Cloud Object Storage [Replay] settings are:

- [All access logs are scanned.]
- All objects that are created or updated since Coordinated Universal Time 00:00:01 from April 11, 2018 to Coordinated Universal Time 10:01:53 on September 21, 2018 are scanned in batches of 1000.
- Custom metadata is retrieved for each object or version.
- Ten vaults are processed in parallel.
- Each vault has a single process LIST that issues requests and 15 processes that issue HEAD requests.

[The following example shows every setting. Most settings have default values and can be omitted., but these screens show a typical example by using default values.]

Example of the Cloud Object Storage Replay settings

```
{
     "system": {
     "name": "Test dsnet",
     "uuid": "00000000-0000-0000-0000-000000000000",
     "manager_ip": "172.1.1.1",
     "accesser_ip": "172.1.1.2",
     "accesser_supports_https": false,
     "manager_username": "admin",
     "manager_password": "password",
     "is_ibm_cos": true
},
   "timestamps": {
     "min_utc": "2018-01-01T00:00:00Z",
     "max_utc": "2018-09-21T10:01:53Z"
},
   “policy_engine” : {
     “spectrum_discover_host”: ”modevvm32.tuc.stglabs.ibm.com”
     “user”: “sdadmin”,
     “password”: “password”
},
   "scanner": {
     "max_requests_per_second": 5000,
     "max_parallel_list": 10,
     "parallel_head_per_list": 5,
     "list_objects_size": 100
},
"notifier":{
     "kafka_format": 1,
     "kafka_endpoint": "192.168.1.1:9092",
     "kafka_topic": "cos-le-connector-topic",
     "kafka_username": "cos",
     "kafka_password": "password",
     "kafka_pem": "-----BEGIN CERTIFICATE-----...\n-----END CERTIFICATE-----\n"
},
   "logging": {
     "debug_log_max_bytes": 10000000,
     "debug_log_backup_count": 10000,
     "notification_log_max_bytes": 10000000,
     "notification_log_backup_count": 10000,
     "notification_log_all": true
},
   "include_all_vaults": false,
   "has_custom_metadata": true,
   "override_warnings": true,
   "exclude-vaults": ["Manager"],
   "vaults": [
     {
       "vault_name": "Vault-1"
     },
     {
       "vault_name": "Vault-2",
       "has_custom_metadata": false
     },
     {
       "vault_name": "Vault-3",
       "has_custom_metadata": false,
       "prefix": "customers/live"
     }
  ]
}
```

Typical Cloud Object Storage configuration settings

```
{
     "dsnet": {
     "name": "Test dsnet",
     "uuid": "00000000-0000-0000-0000-000000000000",
     "manager_ip": "172.1.1.1",
     "accesser_ip": "172.1.1.2",
     "accesser_supports_https": false,
     "manager_username": "admin",
     "manager_password": "password",
     "is_ibm_cos": true
},
   "timestamps": {
     "min_utc": "2018-01-01T00:00:00Z",
     "max_utc": "2018-09-21T10:01:53Z"
},
   “policy_engine” : {
     “spectrum_discover_host”: ”modevvm32.tuc.stglabs.ibm.com”
     “user”: “sdadmin”,
     “password”: “password”
},
   "scanner": {
     "max_requests_per_second": 5000,
     "max_parallel_list": 10,
     "parallel_head_per_list": 5,
     "list_objects_size": 100
},
   "notifier":{
     "kafka_format": 1,
     "kafka_endpoint": "192.168.1.1:9092",
     "kafka_topic": "cos-le-connector-topic",
     "kafka_username": "cos",
     "kafka_password": "password",
     "kafka_pem": "-----BEGIN CERTIFICATE-----...\n-----END CERTIFICATE-----\n"
},
   "logging": {
     "debug_log_max_bytes": 10000000,
     "debug_log_backup_count": 10000,
     "notification_log_max_bytes": 10000000,
     "notification_log_backup_count": 10000,
     "notification_log_all": true
},
   "include_all_vaults": false,
   "has_custom_metadata": true,
   "override_warnings": true,
   "exclude-vaults": ["Manager"],
   "vaults": [
     {
       "vault_name": "Vault-1"
     },
     {
       "vault_name": "Vault-2",
       "has_custom_metadata": false
     },
     {
       "vault_name": "Vault-3",
       "has_custom_metadata": false,
       "prefix": "customers/live"
     }
  ]
}
```

```
{
     "dsnet": {
     "manager_ip": "192.168.2.106",
     "accesser_ip": "192.168.2.111"
},
     "timestamps": {
     "min_utc": "2018-04-11T00:00:01.000Z",
     "max_utc": "2018-09-21T10:01:53Z"
},
     "scanner":{
        "max_requests_per_second": 5000
},
     "include_all_vaults": true
}
```

```
{
     "system": {
     "manager_ip": "192.168.2.106",
     "accesser_ip": "192.168.2.111"
},
    “policy_engine” : {
     “spectrum_discover_host”: ”modevvm32.tuc.stglabs.ibm.com”
},
     "timestamps": {
     "min_utc": "2018-04-11T00:00:01.000Z",
     "max_utc": "2018-09-21T10:01:53Z"
},
     "scanner":{
        "max_requests_per_second": 5000
},
     "include_all_vaults": true
}
```

IBM Cloud Object Storage Scanner is highly configurable. Each element in the file is described in Table 20 on page 102.

Table 20. Explanation of the configuration file

Element	Description	Optional	Default value	Restart scanner if changed	Restart notifier if changed
system section					
`name`	Free-text name of the dsNet. Appears in the `'system_name'` element in all Kafka messages.	√	Retrieved from Manager API if configured. If not the name does not appear in Kafka messages.	√	✗
`uuid`	UUID of the dsNet. Appears in the `'system_uuid'` element in all Kafka messages.	√	Retrieved from Manager API.	√	✗
`manager_ip`	Single IP address or host name of the manager device.	✗	Not applicable	√	✗
`accessor_ip`	Single IP address or host name of an accessor device or load balancer that routes to the accessors.	✗	Not applicable	√	✗
`accessor_supports_https`	Boolean value that indicates whether http or https can be used when you issue requests to the accessor or load balancer.	√	true	√	✗
`manager_username`	User name for accessing the manager API. For testing only. Not to be used in production.	√	Supplied by user at prompt	√	✗
`manager_password`	Password for accessing the Manager API. For testing only. Not to be used in production.	√	Supplied by user at prompt	√	✗

Table 20. Explanation of the configuration file (continued)

Element	Description	Optional	Default value	Restart scanner if changed	Restart notifier if changed
[system section]					
`is_ibm_cos`	Boolean value that indicates whether the system is an IBM Cloud Object Storage or another s3 compliant system. If true, the IBM Get Bucket Extension is used to retrieve object keys from the vaults. **Note:** Setting the value to false is not currently supported by the Scanner and Notifier.	√	True	√	✗
`accessor_access_key`	Access key ID for S3 calls to the accesses or load balancer. For testing only. Not to be used in production.	√	Supplied by user at prompt if you cannot retrieve it from Manager API for the user account that is specified in dsNet/manager_ user name.	√	✗
`accessor_secret_key`	Secret key for S3 calls to the accessor or load balancer. For testing only. Not to be used in production.	√	Supplied by user at prompt if you cannot retrieve from Manager API.	√	✗
Time stamps section					
`min_utc`	Only objects or version in the vaults that have a `LastModified` datetime on or after this time stamp is submitted to IBM Spectrum Discover. Needs to be less than `max_utc`. **Note:** Changing `min_utc` and restarting scanner applies only to objects not yet scanned. Objects scanned before restart might have a `LastModifiedDate` earlier than the new `min_utc`.	✗		√ See note.	✗

Table 20. Explanation of the configuration file (continued)

Element	Description	Optional	Default value	Restart scanner if changed	Restart notifier if changed
[system section]					
max_utc	Only objects or version in the vaults that have a `LastModified` datetime on or before this time stamp is submitted to IBM Spectrum Discover. Needs to be more than min_utc and less than current time. **Note:** Changing `max_utc` to a more recent time and restarting does not mean that new objects written since the old `max_utc` is scanned. The scanner continues from the last object's key scanned in lexicographic order hence new objects with names "less" than the last object scanned is not scanned.	√		√ See note.	✗
Policy engine section		(Only required for IBM Spectrum Discover release 2.0.0.3 and later)			
`spectrum_discover_host`	Host name or IP address of the policy engine service from which the Kafka certificate is retrieved.	✗	none	√	√
`user`	Username for authorization on policy engine.	✗	none	√	√
`password`	Password for authorization on policy engine.	✗	none	√	√
Replay section					
`access_log_directory`	The `access_log_directory` is where the dsNet access log files are stored after download. Access logs must be in the root input folder. Files in subdirectories are not processed.	√	[COS [Replay]]/ access_logs	Restart Replay if changed	Restart Replay if changed

Table 20. Explanation of the configuration file (continued)

Element	Description	Optional	Default value	Restart scanner if changed	Restart notifier if changed
system section					
`download`	If download is set to false, access logs are not downloaded and are assumed to already be present in access_log_directory.	√	true	Restart Replay if changed	Restart Replay if changed
Notifier section		√			
`kafka_format`	Format of the Kafka message.	√	1	✗	√
`kafka_endpoint`	IP address and port of the Kafka endpoint.	√	Retrieved from Manager API	✗	√
`kafka_topic`	Name of the Kafka topic.	√	Retrieved from Manager API	✗	√
`kafka_username`	The user name for authentication with Kafka. **Note:** For testing only. Not to be used in production.	√	Supplied by user at prompt if you cannot retrieve from Manager API.	✗	√
`kafka_password`	The password for authentication with Kafka. **Note:** For testing only. Not to be used in production.	√	Supplied by user at prompt if cannot be retrieved from Manager API.	✗	√
`kafka_pem`	The certificate PEM for authentication with Kafka. Must include '\n' characters to ensure correct formatting. **Note:** For testing only. Not to be used in production.	√	Supplied by user at prompt if it cannot be retrieved from the system	✗	√
Logging section					
`debug_log_max_bytes`	The `scanner.debug` and `notifier.debug` roll over when this size is reached.	√	1,000,000	√	√
`debug_log_backup_count`	The number of `scanner.debug` and `notifier.debug` files to retain.	√	10	√	√
`notification_log_max_b`	The `notification.log` rolls over when this size is reached.	√	1,000,000	√	√
`notification_log_backup_count`	The number of `notification.log` files to retain.	√	10	√	√

Table 20. Explanation of the configuration file (continued)

Element	Description	Optional	Default value	Restart scanner if changed	Restart notifier if changed
system section					
`notification_log_all`	Boolean value that controls the level of Notifier logging. When true: an entry is written to `notification.log` for message you send to the Kafka cluster. When false: only failed sends are written to `notification.log`.	√	False	✗	√
Root-level items					
`include_all_vaults`	Boolean value that determines whether all vaults in the dsNet are scanned. If false, the details of the vaults to be scanned must be specified in the 'vaults' element. Boolean value that determines whether custom metadata and content type are retrieved for each object by using individual HEAD requests.	√	False	√	✗
`has_custom_metadata`	This value is only relevant when a versioned vault is scanned. For IBM Cloud Object Storage systems, non-versioned vaults always require a HEAD request for every object. Can be overridden for each vault in the 'vaults' element.	√	True	√	✗
`override_warnings`	Boolean value that allows the scanner to run and ignore any warnings that are generated on start-up. For example, a warning is raised on start-up if versioning is suspended on a vault.	√	False	√	✗
`exclude_vaults`	Comma-separated list of vault names to be excluded from scanning. For example: `"exclude-vaults": ["COSVault", "COSVault-V"]`	√	[] Empty list	√	✗

Table 20. Explanation of the configuration file (continued)

Element	Description	Optional	Default value	Restart scanner if changed	Restart notifier if changed
[system section]					
`vaults`	List of vaults to be scanned. If `include_all_vaults` is true the vaults list can be left empty. This list can be used to define more detailed scanning parameters for individual vaults. Any settings that are defined here take precedence over the settings that are described previously. Each element in the list contains: The `vault_name` is the name of the vault. The `has_custom_metadata` is an optional boolean that overrides the `has_custom_metadata` that is described. The `prefix` is an optional string that is used to filter the objects or versions that are retrieved from the vault.	√	Dependent on settings include_all_vault s and exclude_vaults	√	✗

[Replay] performance

The number of requests that are issued by IBM Cloud Object Storage [Replay] is throttled to ensure that overall dsNet performance remains at the agreed level.

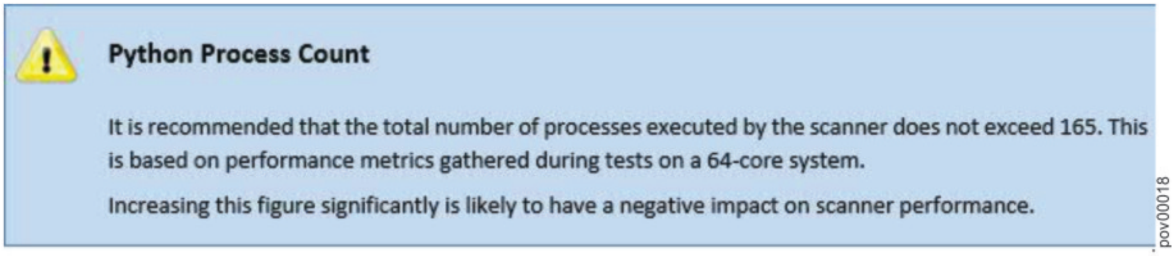
You can control throttling by the number of settings in a configuration file. All settings are optional. The following screen shows an example of the default values.

```
"[replay]": {
     "max_requests_per_second": 1000,
     "max_parallel_list": 10,
     "parallel_head_per_list": 15,
     "list_objects_size": 1000
}
```

Process count

The following list shows an example of how 161 processes are divided. Figure 20 on page 108 shows a caution message of how the number of processes should not exceed 161.

- One main process
- 10 List worker processes
- 150 HEAD worker processes



Python Process Count

It is recommended that the total number of processes executed by the scanner does not exceed 165. This is based on performance metrics gathered during tests on a 64-core system.

Increasing this figure significantly is likely to have a negative impact on scanner performance.

Figure 20. Python process count

Maximum Replay performance

Replay and Notifier maximize performance on a 64 core Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz server is 2300 objects that are scanned and notified per second with a dsNet with 6 accessors and 12 slicestors under customer load at 50 percent capacity.

```
"replay": {
     "max_requests_per_second": 2300,
     "max_parallel_list": 10,
     "parallel_head_per_list": 15,
     "list_objects_size": 1000
 }
```

The recommendation is to start the replay at a rate of 1000 objects scanned per second. Measure the latency degradation of customer traffic and increase the scanning rate until the maximum acceptable degradation is reached.

One thousand objects per second on the net, which is a 5 - 27 percent increase of write operations, the latency (larger increase for smaller size files) and around 10 percent for read operations latency were measured.

At 2000 objects a second, a 10 - 50 percent increase of write operations latency and in the range 18 - 28 percent, and 10 percent for read operations latency were measured.

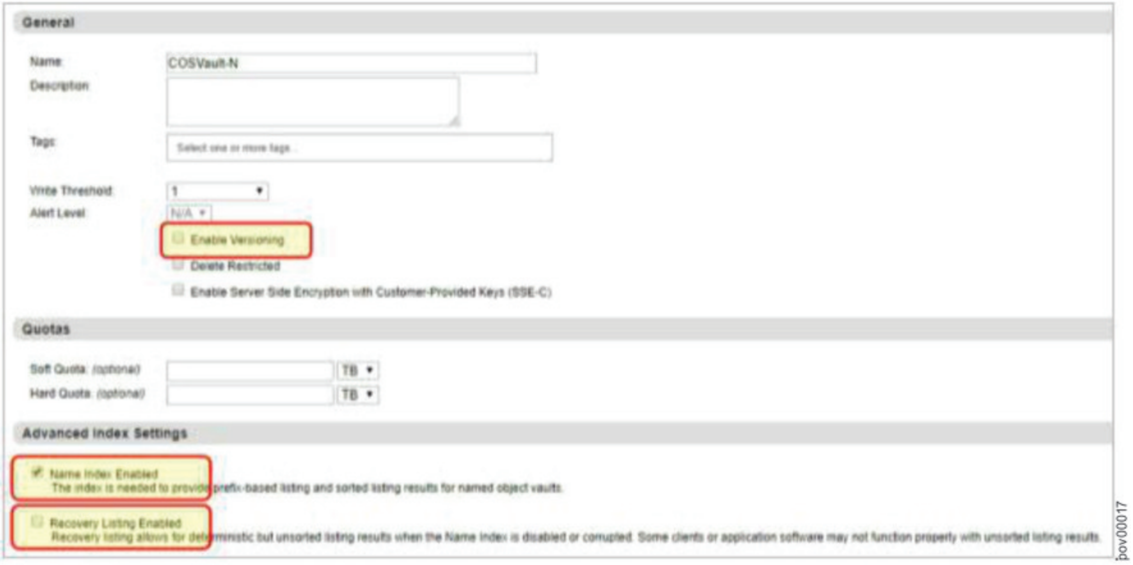
Replay tasks and vault settings

A few scenarios exist that prevent the Replay from operating correctly.

Certain combinations of the following IBM COS vault settings prevent the Replay from executing a full scan:

- Vault versioning
- Name index
- Recover listing

Figure 21 on page 109 shows settings for the first three items on the vault configuration page in the DsNet Manager user interface.



General

Name: COSVault-N

Description:

Tags: Select one or more tags...

Write Threshold: 1 ▾

Alert Level: N/A ▾

□ Enable Versioning

□ Delete Restricted

□ Enable Server Side Encryption with Customer-Provided Keys (SSE-C)

Quotas

Soft Quota: *(optional)* TB ▾

Hard Quota: *(optional)* TB ▾

Advanced Index Settings

☑ Name Index Enabled
The index is needed to provide prefix-based listing and sorted listing results for named object vaults.

□ Recovery Listing Enabled
Recovery listing allows for deterministic but unsorted listing results when the Name Index is disabled or corrupted. Some clients or application software may not function properly with unsorted listing results.

Figure 21. Settings for three items on the vault configuration page in the net Manager user interface

The scenarios that are invalid are reported at startup.

Remember: You must correct the scenarios before you can run the Replay.

If a scan does not complete successfully, make sure that you check the log file for errors and warnings. In some cases, you must modify the settings file as detailed in the errors and warning messages. The settings file is located at: `/gpfs/gpfs0/connections/cos/scan/scanner-settings.json`

Table 21 on page 109 shows the behavior for the Replay for different combinations of the four variables.

Table 21. Behaviors for Replay for four variables

ID	Name index	Recovery listing	Versioned	Cloud Object Storage Replay behavior
0	✗	✗	✗	□ Stop start-up and report error in config file: Error: Objects cannot be listed because **Name Index** and **Recovery Index** are both disabled. You might enable **Recovery Listing** on the vault or add this vault to the `"exclude_vaults"` list in the configuration file. For example: `"exclude-vaults": ["vault-name"]"`
1	✗	✗	√	□ Stop start-up and report error in config file: Error: Objects cannot be listed because **Name Index** and **Recovery Index** are both unavailable. You might enable **Recovery Listing** on the vault or add this vault to the `"exclude_vaults"` list in the configuration file. For example: `"exclude-vaults": ["vault-name"]"`

Table 21. Behaviors for Replay for four variables (continued)

ID	Name index	Recovery listing	Versioned	Cloud Object Storage Replay behavior
2	✗	√	✗	☑ Object Listing is run.
3	✗	√	√	☑ Object Listing runs. Only the most recent version of each object will be listed. A warning is logged: Warning: Versions cannot be listed as Name Index is unavailable. An object scan will be executed and only the most recent version of each object is listed. You must add override_warnings: true in the config file to ignore this warning. Switching Name Index on will not enable scanning of a full version history. Objects created while Name Index is off will not be present when it is enabled.
4	√	✗	✗	☑ Object Listing will be executed
5	√	✗	√	☑ Object Listing is run.
6	√	√	✗	☑ Object Listing is run.
7	√	√	√	! Stop start-up and report warning: This is a versioned vault but version scanning is not possible as **Recovery Listing** is enabled. You might either disable **Recovery Listing** on the vault to allow version scanning, or rerun the Replay with the argument `override-warnings: true` to allow object scanning.

]

Important: [You might receive system errors about records not being scanned to the database if you scan a COS vault with **Name Index** disabled and **Recovery Listing** enabled. You cannot specify a prefix listing on a vault that has **Recovery Listing** enabled (you cannot specify a blank prefix either).]

Including and excluding vaults

You can set the vaults that you scan with various settings in the configuration file.

Use the following settings in the configuration file to scan the vaults:

- `include_all_vaults` (Boolean)
- `exclude_vaults` (List)
- `vaults` (Dice)

When `include_all_vaults` is true, all vaults in the system are scanned except for any vaults specified in the `exclude_vaults` list.

You might consider `exclude_vaults` a blacklist of vaults to ignore and `vaults` is a whitelist that specifies details of individual vaults to be scanned.

If `include_all_vaults` is true and the vaults list is populated, the list of vaults that are scanned is the superset of all vaults that are returned by the Manager that are merged with the vaults list from the config file.

An error is raised and the Scanner aborts on start-up if the same vault appears in both `vaults` and `exclude_vaults`.

Mirror, Proxy, Data Migration

IBM Cloud Object Storage Scanner does not support scanning of the following:

- Mirrored vaults
- Proxy vaults
- Vaults that are set up for migration

Any vaults of these types are ignored by the scanner and a warning logged in the debug log.

Examples for including and excluding vaults

To summarize the rules for including and excluding vaults, following are some examples:

Example 1

- The system contains 1000 vaults.
- Five of the 1000 vaults are management vaults (named mgmt-1 to mgmt-5)
- The scan includes all vaults except the management vaults

```
"include_all_vaults": true,
"exclude-vaults": ["mgmt-1", "mgmt-2", "mgmt-3", "mgmt-4", "mgmt-5"]
```

Example 2

- The system contains 1000 vaults.
- 5 of the 1000 vaults are management vaults (named mgmt-1 to mgmt-5)
- The scan includes all vaults except the management vaults
- The scan includes a filter for scanning a vault that is named vault-x.
- The scan includes only a scan of the objects whose key starts with **production/finance**.

```
"include_all_vaults": true,
"exclude-vaults": ["mgmt-1", "mgmt-2", "mgmt-3", "mgmt-4", "mgmt-5"],
"vaults": [
     {"vault_name":"vault-x", "prefix":"production/finance"}
  ]
```

Example 3

- The system contains 1000 vaults.
- 5 of the 1000 vaults are management vaults (named mgmt-1 to mgmt-5)
- The scan includes all vaults except the management vaults
- The scan includes a filter for scanning a vault that is named vault-x.
- The scan includes only a scan of the objects whose key starts with **production/finance** or **production/marketing**.

```
"include_all_vaults": true,
"exclude-vaults": ["mgmt-1", "mgmt-2", "mgmt-3", "mgmt-4", "mgmt-5"],
"vaults": [
     {"vault_name":"vault-x", "prefix":"production/finance"},
     {"vault_name":"vault-x", "prefix":"production/marketing"}
   ]
```

Example 4

- The system contains 1000 vaults.
- Run a test on three vaults named vault-a, vault-b, and versioned-vault-c.
- Run a scan on versioned-vault-c and issue LIST requests. Do not issue HEAD requests because the objects do not have custom amz headers.

```
"include_all_vaults": false,
"vaults": [
     {"vault_name":"vault-a"
     {"vault_name":"vault-b"
     {"vault_name":"vault-c",  "has_custom_metadata":false}
   ]
```

Stats files
The IBM Cloud Object Storage Scanner tracks each LIST process status to a stats file.

[During a scan, the Scanner runs multiple processes. Each `LIST` processes and tracks the progress, saves the `next_key`, and optionally the `next_version` to a stats file named `task.stats` that is stored with the log files in the `/gpfs/gpfs0/connections/cos/replay/output/data` directory.]

```
{
     "estimated_object_count": 1000,
     "list_objects_size": 100,
      "next_key": "",
      "next_version": "",
      "prefix": "",
      "scan_type": "Object Scan",
      "status": "Complete",
      "total_bytes_output": 1126809,
      "total_bytes_scanned": 1126809,
      "total_objects_output": 47,
      "total_objects_scanned": 47,
      "vault_name": "dsmgmt-sp1",
      "vault_uuid": "868daa21-9e56-4c41-b6fd-845a4c85cea9"
}
```

From the Scanner, you can start, stop, recover files from a crash, and restart at the point where the scan was interrupted.

When you start the scanner:

1. Processing of the Scanner continues from `next_key` and `next_version`.
2. Queue of the Notifier is optimized by reloading from the files in the data folder instead of re-querying the dsNet.
3. Batches that were processed partially are reprocessed. Duplicate Kafka notifications might occur, but are handled safely by the IBM Spectrum Discover system.

Replay
When a severe outage occurs and causes the loss of notifications sent by the system to IBM Spectrum Discover, the IBM Cloud Object Storage Scanner Replay feature can be used to recover lost notifications.

Replay parses the access logs of a system and reconstitutes the notifications. Also, the Notifier can resend the notifications.

Initialization for Replay
During the startup, Replay reads the configuration file and issues requests to the Manager of the dsNet device similar to the Scanner.

Data from the configuration file is validated to ensure that appropriate permissions are granted in the dsNet. This allows access to management vaults and regular vaults. Startup errors or warnings are logged and printed to the console.

After initialization, Replay extracts accessor log files from the management vaults of dsNet and enables Replay to process and write notifications to the output directory.

Error conditions

Sometimes Replay does not have enough information to replay the original notification. If this occurs, you must fix the problems manually.

For example, if vault versioning was suspended when you made the request and you receive an s3 DeleteObject for an object or delete marker, the following error is logged:

```
error_code=True, error_description=”Delete operation with [no version_id|null
version_id|version_id] for vault with versioning = [suspended/enabled]”
```

The error message displays because Replay cannot distinguish when a notification with s3:CreateDeleteMarker or s3:CreateDeleteMarker:NullVersionDeleted is sent.

If vault versioning is disabled, and an s3 PutObject request is received for an object that is deleted, the following error is logged:

```
error_code=404, error_message=”Not Found”
```

The error message displays because Replay cannot determine the tag of the object that was deleted.

Output

Messages are batched by 1,000 or to the Scanner list objects size configuration setting, if specified.

The messages are written to the output folder with the same Notification format used by the Scanner.

```
{
     "system_name": "Test",
     "object_etag": "\"de37d2cee49596916f62a233dfc790a4\"",
     "request_time": "2018-09-24T18:49:29.383Z",
     "format": 1,
     "bucket_uuid": "ac89915b-d4ec-7ff1-00be-9c32b2aca580",
     "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865",
     "object_length": "12319",
     "object_name": "test_version",
     "bucket_name": "vault3",
     "content_type": "binary/octet-stream",
     "request_id": "17451c3d-e81e-40ed-939a-4534780daaa8",
     "operation": "s3:PutObject"
}
```

If an error occurs, the error messages are written to the `/gpfs/gpfs0/connections/cos/replay/output/data/access_log_error/` directory. Take note of the extra `error_code` and `error_description` elements.

```
{
     "system_name": "Test",
     "object_version": "null",
     "request_time": "2018-09-24T17:07:59.471Z",
     "format": 1,
     "bucket_uuid": "ac89915b-d4ec-7ff1-00be-9c32b2aca580",
     "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865",
     "object_length": "12319",
     "object_name": "object5.2",
     "bucket_name": "vault3",
     "request_id": "ebc472a1-f955-4605-895b-840867b12e01",
     "operation": "s3:PutObject",
     "error_description": "Not Found",
     "error_code": 404
}
```

Renaming a vault for Replay

When you rename a vault, it is possible that Replay can abort.

Replay aborts when you:

- Delete the vault.
- Rename the vault.

- Discover that the read permission is revoked for the credentials that are supplied by the operator or manager API.

All other scans of a vault continue scanning until complete.

[You can find the details of the errors that include stack trace in the `replay.debug` file in the `/gpfs/gpfs0/connections/cos/replay/debug/replay/[timestamp]` directory.]

Starting the Replay

The guidelines and rules for using Replay are documented in this topic.

To start Replay, run the following command:

```
[cos-replay]
```

The following rules apply for Replay:

- Configure Replay according to the guidelines in Table 20 on page 102.
- Replay component requires `min_utc` and `max_utc` time stamps defined in the "Configuration file" on page 99.
- Only notifications sent between `min_utc` and `max_utc` are parsed and replayed.
- Replay automatically shuts down when all accessor logs are downloaded and processed. The message **Complete Replay Process** appears in the console.

This is an example of how to start Replay:

```
Starting COS Replay - Version 0.1 Log file and config file are in directory /Users/weebrew/
Documents/Development/ibmworkspace/cosscanner/output/debug/scanner/20180925-125528-232131
Starting Accessor Log Extraction Downloading files...
('Downloaded', 10, 'of', 36)
('Downloaded', 20, 'of', 36)
('Downloaded', 30, 'of', 36)
Download complete.
Total files: 36
Complete Accessor Log Extraction
Starting Replay process
Complete Replay process
```

Debug mode for Replay

Run Replay in debug mode to troubleshoot problems.

To start debug mode, run the following command:

```
cos-replay --log=DEBUG
```

Running debug mode creates large log files and creates a significant drop in performance. Do not run debug mode for long periods especially when you are in production mode.

Notifier

The Notifier is the component that reads the JSON notifications that are written by Scanner or Replay and sends notifications to the Kafka cluster.

When notifications are acknowledged by Kafka, the Notifier moves the file to the archive folder.

On start-up, Notifier calls the Manager API and retrieves details of any Notification Service Configurations (NSC) configured in the dsNet for IBM Spectrum Discover. If more than one is found, the first one is used.

Retrieval of NSCs is overridden by defining the details of the Kafka configuration in the config file.

```
 "notifier":{
   "kafka_format": 1,
   "kafka_endpoint": "192.168.1.34:9092",
   "kafka_topic": "cos-le-connector-topic"
}
```

Limitations

Limitations apply when the Notifier uses a Kafka configuration retrieved from the Manager API.

- If more than one NSC exists, the first one is used for all vaults.
- If more than one host name is defined in the NSC, the first one is used for all vaults.

Starting the Notifier

Running the Notifier has rules and limitations.

To start the Notifier, run the following command:

```
cos-notify
```

After you start the Notifier, you are prompted for security credentials for the manager API and Kafka cluster.

```
Starting COS Notifier - Version 0.1
Enter the Manager API username: admin
Enter the Manager API password:
Enter the Kafka username: cos
Enter the Kafka password:
Enter the Kafka pem:
Creating Kafak producer...
Done
Notifier is running
Log file and config file are in directory C:\dev\cos-scanner\output\debug\notifier
\20180912-121641-283000
Checking for files in   \data
- 11 files found Checking for files in output\data
- 256 files found
```

Rules and limitations

The following rules and limitations apply to the Notifier:

- You cannot start the Notifier in the background because the Notifier requires user input at the terminal window.
- You can stop the Notifier and force the Notifier to run in the background.
- The passwords and pem do not display when you type and paste the passwords in the console.
- The certificate pem is approximately 1,600 characters. If you use an SSH connection, the certificate pem might be truncated to 1,000 characters.
- If the number is truncated to 1,000 characters, include the certificate pem in the config file.

Notifier operation

The Notifier enumerates and processes all .log files in the Scanner data directory.

After all files are processed, the Notifier repeats the process so that new `.log` files that are generated by the Scanner are processed. The Notifier sleeps repeatedly in 1-second intervals if no new files are found in the `/gpfs/gpfs0/connections/cos/replay/output/data` directory.

The Notifier does not automatically shut down. The Notifier continues to monitor the Scanner data directory for new .log files. Monitor the progress of the Notifier by using the status report. When the operator or administrator determines that all scanned objects are submitted successfully to the IBM Spectrum Discover, shut down the Notifier by using the kill switch.

Stopping the Notifier

You might need to stop and restart the Notifier.

Before you stop and restart the Notifier:

1. Create a file named `kill.notifier` in the `/gpfs/gpfs0/connections/cos/replay/output/command` directory.

2. Ensure that the processing of any batches is complete before you stop the Notifier.

Stopping the Notifier displays the following output:

The shutdown is complete when the "**`Shutdown is complete`**" message displays.

```
Starting COS Notifier - Version 0.1
Enter the Manager API username: admin
Enter the Manager API password:
Enter the Kafka username: cos Enter the Kafka password:
Enter the Kafka pem: Creating Kafak producer...
Done
Notifier is running
Log file and config file are in directory C:\dev\cos-scanner\output\
   debug\notifier\20180912-121641-283000
Checking for files in output\data
- 11 files found Checking for files in output\data
- 256 files found Detected the kill trigger file. Shutting down...
Shutdown is complete
```

Restarting the Notifier

When you stop the Notifier following a shutdown with the kill.notifier file, you must rename the file manually or delete the file before you do a restart.

If you do not rename or delete the kill.notifier file, the system finds the file and displays the following message:

```
C:\dev\cos-scanner>python main_notifier.py Starting COS Notifier - Version 0.1
The file 'kill.notifier' is preventing the notifier from running.
You should delete or rename the file and re-start the notifier.
File location: 'C:\dev\cos-scanner\command\kill.notifier
```

The Scanner and Notifier are separate solutions that share the config file. The files operate independently, so you can start and stop either file independently any time.

Progress report

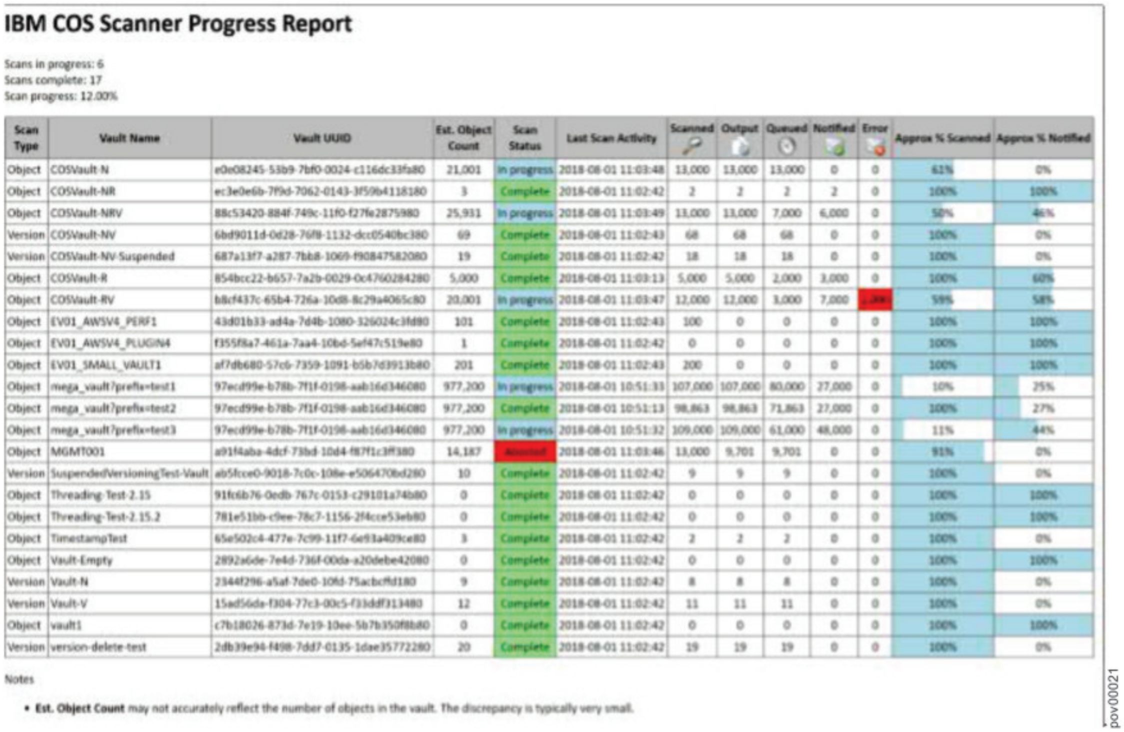
The Progress Report provides an instant snapshot of status for the Scanner and Notifier.

To create a progress report, run the following command:

```
cos-report
```

The progress report displays in plain text format to the console in a static HTML file that is named: /`gpfs/gpfs0/connections/cos/replay/output/cos-scanner-report.html`

If a progress report exists, the new progress report overwrites the existing progress report. See Figure 22 on page 117.



IBM COS Scanner Progress Report

Scans in progress: 6
Scans complete: 17
Scan progress: 12.00%

Scan Type	Vault Name	Vault UUID	Est. Object Count	Scan Status	Last Scan Activity	Scanned	Output	Queued	Notified	Error	Approx % Scanned	Approx % Notified
Object	COSVault-N	e0e08245-53b9-7bf0-0024-c116dc33fa80	21,001	In progress	2018-08-01 11:03:48	13,000	13,000	13,000	0	0	61%	0%
Object	COSVault-NR	ec3e0e6b-7f9d-7062-0143-3f59b4118180	3	Complete	2018-08-01 11:02:42	2	2	2	2	0	100%	100%
Object	COSVault-NRV	88c53420-884f-749c-11f0-f27fe2875980	25,931	In progress	2018-08-01 11:03:49	13,000	13,000	7,000	6,000	0	50%	46%
Version	COSVault-NV	6bd9011d-0d28-76f8-1132-dcc0540bc380	69	Complete	2018-08-01 11:02:43	68	68	68	0	0	100%	0%
Version	COSVault-NV-Suspended	687a13f7-a287-7bb8-1069-f90847582080	19	Complete	2018-08-01 11:02:42	18	18	18	0	0	100%	0%
Object	COSVault-R	854bcc22-b657-7a2b-0029-0c4760284280	5,000	Complete	2018-08-01 11:03:13	5,000	5,000	2,000	3,000	0	100%	60%
Object	COSVault-RV	b8cf437c-65b4-726a-10d8-8c29a4065c80	20,001	In progress	2018-08-01 11:03:47	12,000	12,000	3,000	7,000	[illegible]	59%	58%
Object	EV01_AWSV4_PERF1	43d01b33-ad4a-7d4b-1080-326024c3fd80	101	Complete	2018-08-01 11:02:43	100	0	0	0	0	100%	100%
Object	EV01_AWSV4_PLUGIN4	f355f8a7-461a-7aa4-10bd-5ef47c519e80	1	Complete	2018-08-01 11:02:42	0	0	0	0	0	100%	100%
Object	EV01_SMALL_VAULT1	af7db680-57c6-7359-1091-b5b7d3913b80	201	Complete	2018-08-01 11:02:43	200	0	0	0	0	100%	100%
Object	mega_vault?prefix=test1	97ecd99e-b78b-7f1f-0198-aab16d346080	977,200	In progress	2018-08-01 10:51:33	107,000	107,000	80,000	27,000	0	10%	25%
Object	mega_vault?prefix=test2	97ecd99e-b78b-7f1f-0198-aab16d346080	977,200	Complete	2018-08-01 10:51:13	98,863	98,863	71,863	27,000	0	100%	27%
Object	mega_vault?prefix=test3	97ecd99e-b78b-7f1f-0198-aab16d346080	977,200	In progress	2018-08-01 10:51:32	109,000	109,000	61,000	48,000	0	11%	44%
Object	MGMT001	a91f4aba-4dcf-73bd-10d4-f87f1c3ff380	14,187	Aborted	2018-08-01 11:03:46	13,000	9,701	9,701	0	0	91%	0%
Version	SuspendedVersioningTest-Vault	ab5fcce0-9018-7c0c-108e-e506470bd280	10	Complete	2018-08-01 11:02:42	9	9	9	0	0	100%	0%
Object	Threading-Test-2.15	91fc6b76-0edb-767c-0153-c29101a74b80	0	Complete	2018-08-01 11:02:42	0	0	0	0	0	100%	100%
Object	Threading-Test-2.15.2	781e51bb-c9ee-78c7-1156-2f4cce53eb80	0	Complete	2018-08-01 11:02:42	0	0	0	0	0	100%	100%
Object	TimestampTest	65e502c4-477e-7c99-11f7-6e93a409ce80	3	Complete	2018-08-01 11:02:42	2	2	2	0	0	100%	0%
Object	Vault-Empty	2892a6de-7e4d-736f-00da-a20debe42080	0	Complete	2018-08-01 11:02:42	0	0	0	0	0	100%	100%
Version	Vault-N	2344f296-a5af-7de0-10fd-75acbcffd180	9	Complete	2018-08-01 11:02:42	8	8	8	0	0	100%	0%
Version	Vault-V	15ad56da-f304-77c3-00c5-f33ddf313480	12	Complete	2018-08-01 11:02:42	11	11	11	0	0	100%	0%
Object	vault1	c7b18026-873d-7e19-10ee-5b7b350f8b80	0	Complete	2018-08-01 11:02:42	0	0	0	0	0	100%	100%
Version	version-delete-test	2db39e94-f498-7dd7-0135-1dae35772280	20	Complete	2018-08-01 11:02:42	19	19	19	0	0	100%	0%

Notes

- **Est. Object Count** may not accurately reflect the number of objects in the vault. The discrepancy is typically very small.

Figure 22. IBM Cloud Object Storage Scanner progress report

See Table 22 on page 117 for a description of information in the progress report.

Table 22. Description for IBM Cloud Object Storage Scanner progress report

Column name	Description
Scan Type	Either Object or Version. Non-Versioned vaults show Object. Versioned vaults show Version. But, some exceptions exist. If the Name Index for a versioned vault is unavailable but Recovery Listing is enabled, an object scan might be run. The user is alerted that an object scan can be done, but this object scan requires changes to the configuration file.
Vault Name	The name of the vault. Any prefix that is defined in the configuration file is also shown. Example: `mega_vault?prefix=test`
Vault UUID	The UUID of the vault.
Estimated Object Count	The estimated number of objects in the vault, as reported by the Manager API. This value is refreshed from the Manager API each time the scanner is started, regardless of the status of each scan. Given that the number of objects in each vault might be constantly changing, the number of objects that are reported in this column becomes out of date during long running scans. **Note:** This issue affects only the Status Report but does not affect the data integrity of the Scanner.

Table 22. Description for IBM Cloud Object Storage Scanner progress report (continued)

Column name	Description
Scan Status	Shows the status of the scanner. **Not started** The task is queued but not started. **In progress** The task is running. **Complete** The task finished. **Aborted** The task encountered an unrecoverable error and aborted. Shut down the Scanner and the debug file, and inspect the file to investigate the problems. After you resolve the problems, restart the Scanner. The debug file is in the `/gpfs/gpfs0/connections/cos/replay/output/data/<vault-name>/<prefix>` directory. For each vault, see the `/gpfs/gpfs0/connections/cos/replay/output/data/<vault-name>/<prefix>` directories.
Last scan activity	The last time data was retrieved from the vault.
Scanned	Number of objects/versions scanned. For a versioned vault, this shows a figure that is higher than the Estimated Object Count.
Output	Number of objects/versions that scanned AND whose LastModified time stamp is inside the time window that is defined in the configuration file. The figure in the column is Queued + Notified + Error.
Queued	Number of objects or versions that are Output and are waiting to be sent to the Kafka cluster.
Notified	Number of objects/versions that are submitted successfully to the Kafka cluster.
Error	Number of objects/versions that failed to send to the Kafka cluster. Details of all errors are logged to notifier.debug.
Approximate percentage scanned	Scanned as a percentage of Est. Object Count. The cell background shows a progress bar.
Approximate percentage scanned	Notified as a percentage of Output. The cell background shows a progress bar.

Table 23. What is reported beneath the report title

Measure	Description
Scans in progress	Number of scans with the status "In Progress". Applies to Scanner only.
Scans complete	Number of scans with the status "Complete". Applies to Scanner only.
Scan progress	Sum (number of objects scanned) as a percentage of sum (estimated object count).

Logging

You can view the list of directories generated by scanner, notifier, and replay.

Table 24 on page 119 lists the directories generated on start-up by the Scanner, Notifier and Replay.

Table 24. List of directories generated by scanner, notifier, and replay

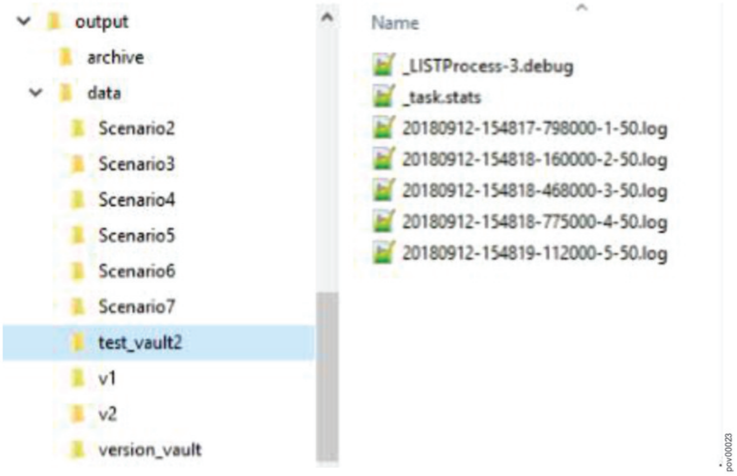
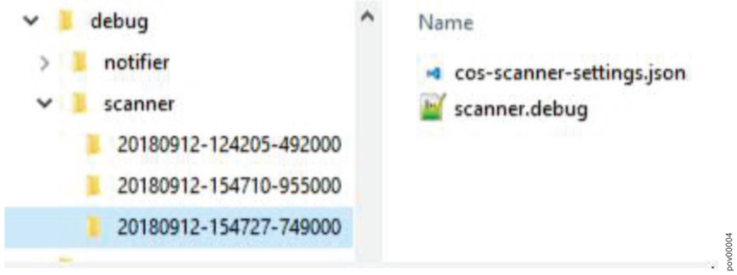
Directory	**Description**
[For IBM Spectrum Discover: `/gpfs/gpfs0/connections/cos/replay/output/data`]	Contains .log files (Kafka messages), stats files and debug information for each scanned vault.  output archive data Scenario2 Scenario3 Scenario4 Scenario5 Scenario6 Scenario7 test_vault2 v1 v2 version_vault Name _LISTProcess-3.debug _task.stats 20180912-154817-798000-1-50.log 20180912-154818-160000-2-50.log 20180912-154818-468000-3-50.log 20180912-154818-775000-4-50.log 20180912-154819-112000-5-50.log pov00023
[For IBM Spectrum Discover: `/gpfs/gpfs0/connections/cos/replay/output/debug/[scanner|replay]/`]	Contains Scanner/Replay debug/troubleshooting information. A new sub-directory is created each time the Scanner/Replay starts. Each sub-directory contains a copy of the configuration file and `scanner.debug` (`replay.debug`).  debug notifier scanner 20180912-124205-492000 20180912-154710-955000 20180912-154727-749000 Name cos-scanner-settings.json scanner.debug pov00004
[For IBM Spectrum Discover: `/gpfs/gpfs0/connections/cos/replay/output/debug/notifier`]	Contains Notifier debug/troubleshooting information. A new sub-directory is created each time the Notifier starts. It contains a copy of the configuration file and `notifier.debug`. Same directory naming convention as shown above for the scanner. `notifier.debug` will rollover when it reaches a predefined size as defined in the configuration file. See "Configuration file" on page 99.
[For IBM Spectrum Discover: `/gpfs/gpfs0/connections/cos/replay/output/archive/`]	Contains all .log files that have been successfully processed by the Notifier and their contents successfully submitted to the Kafka cluster. Note that files in this directory are truncated – they contain only the object key and (optionally) the version.
[For IBM Spectrum Discover: `/gpfs/gpfs0/connections/cos/replay/output/error/`]	Contains any .log files that failed to submit to the Kafka cluster.

Table 24. List of directories generated by scanner, notifier, and replay (continued)

Directory	**Description**
[For IBM Spectrum Discover: `/gpfs/gpfs0/connections/cos/replay/output/notification_log/`]	The `notification.log` file contains details of any errors (including stack trace) that occurred when attempting to send notifications to the Kafka cluster. If `logging/notification_log_all` is true in the config file, all successful sends are also logged. The `notification.log` file will rollover when it reaches a predefined size as defined in the configuration file. See "Configuration file" on page 99.

IBM Cloud Object Storage Scanner output data

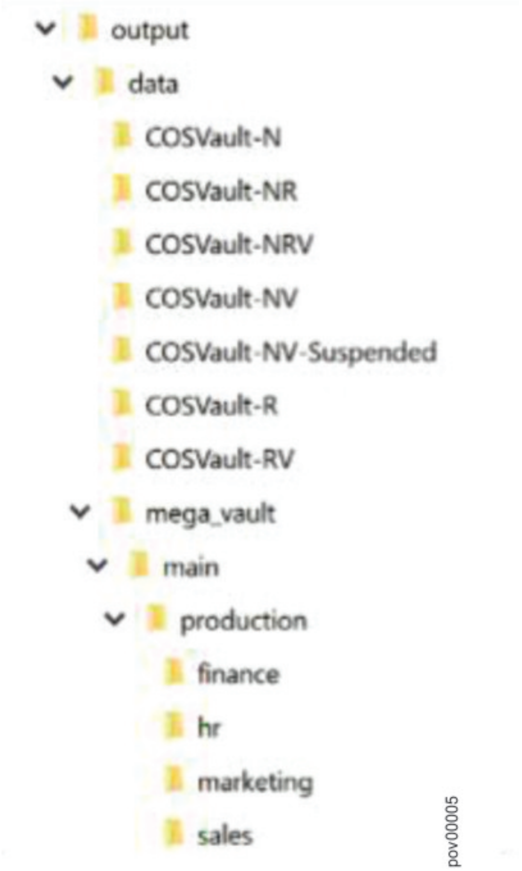
The Scanner generates a directory beneath the output data directory for each vault or vault prefix as defined in the configuration file.

The `/gpfs/gpfs0/connections/cos/replay/output/data` directory is the Scanner output data directory.

The following screen shows an example of a configuration file and also shows that all vaults are scanned, but mega_vault has four separate prefixes that are defined winch means the four scans of the vault occurred.

```
"include_all_vaults": true,
  "vaults": [
  {"vault_name": "mega_vault", "prefix": "main/production/finance"},
  {"vault_name": "mega_vault", "prefix": "main/production/sales"},
  {"vault_name": "mega_vault", "prefix": "main/production/marketing"},
  {"vault_name": "mega_vault", "prefix": "main/production/hr"}
]
```

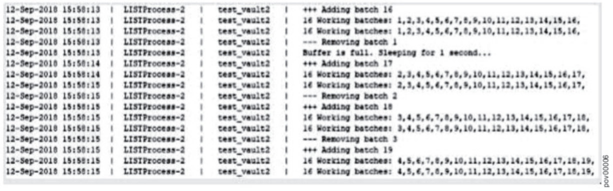
Figure 23 on page 121 shows the directory structure.



```
output
  data
    COSVault-N
    COSVault-NR
    COSVault-NRV
    COSVault-NV
    COSVault-NV-Suspended
    COSVault-R
    COSVault-RV
  mega_vault
    main
      production
        finance
        hr
        marketing
        sales
```

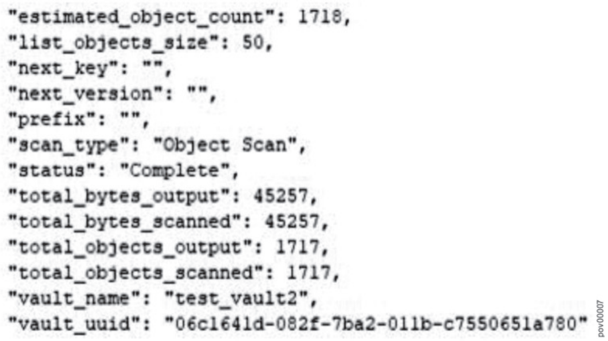
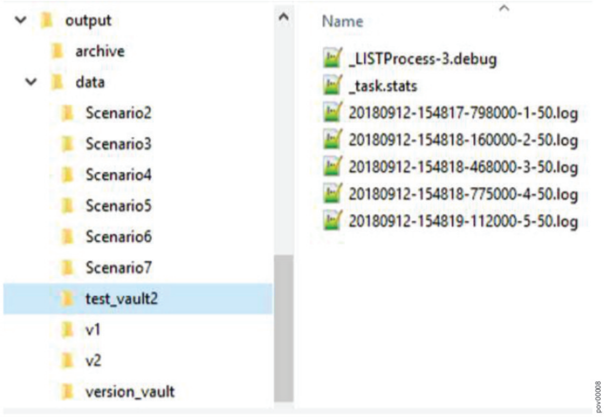
Figure 23. Directory structure from the configuration file

The status and progress of each scan must be maintained so a separate directory structure is created for each scan. Table 25 on page 121 shows the leaf directories that contain the file names and description.

<table>
<tr><td colspan="2"><em>Table 25. Leaf directory file names</em></td></tr>
<tr><th>File name</th><th>Description</th></tr>
<tr><td>_LISTProcessN.debug</td><td>The N in the file name is different for each process (0 - 9 if there are 10 processes).
Contains detailed debug information and details of any errors that are encountered when you scan the vault. Figure 24 on page 121 shows an example of running in debug mode.

<pre>
12-Sep-2018 15:58:13 | LISTProcess-2 | test_vault2 | +++ Adding batch 16
12-Sep-2018 15:58:13 | LISTProcess-2 | test_vault2 | 16 Working batches: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,
12-Sep-2018 15:58:13 | LISTProcess-2 | test_vault2 | 16 Working batches: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,
12-Sep-2018 15:58:13 | LISTProcess-2 | test_vault2 | --- Removing batch 1
12-Sep-2018 15:58:13 | LISTProcess-2 | test_vault2 | Buffer is full. Sleeping for 1 second...
12-Sep-2018 15:58:14 | LISTProcess-2 | test_vault2 | +++ Adding batch 17
12-Sep-2018 15:58:14 | LISTProcess-2 | test_vault2 | 16 Working batches: 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | 16 Working batches: 2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | --- Removing batch 2
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | +++ Adding batch 18
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | 16 Working batches: 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | 16 Working batches: 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | --- Removing batch 3
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | +++ Adding batch 19
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | 16 Working batches: 4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,
12-Sep-2018 15:58:15 | LISTProcess-2 | test_vault2 | 16 Working batches: 4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,
</pre>

<em>Figure 24. Example of running in debug mode</em></td></tr>
</table>

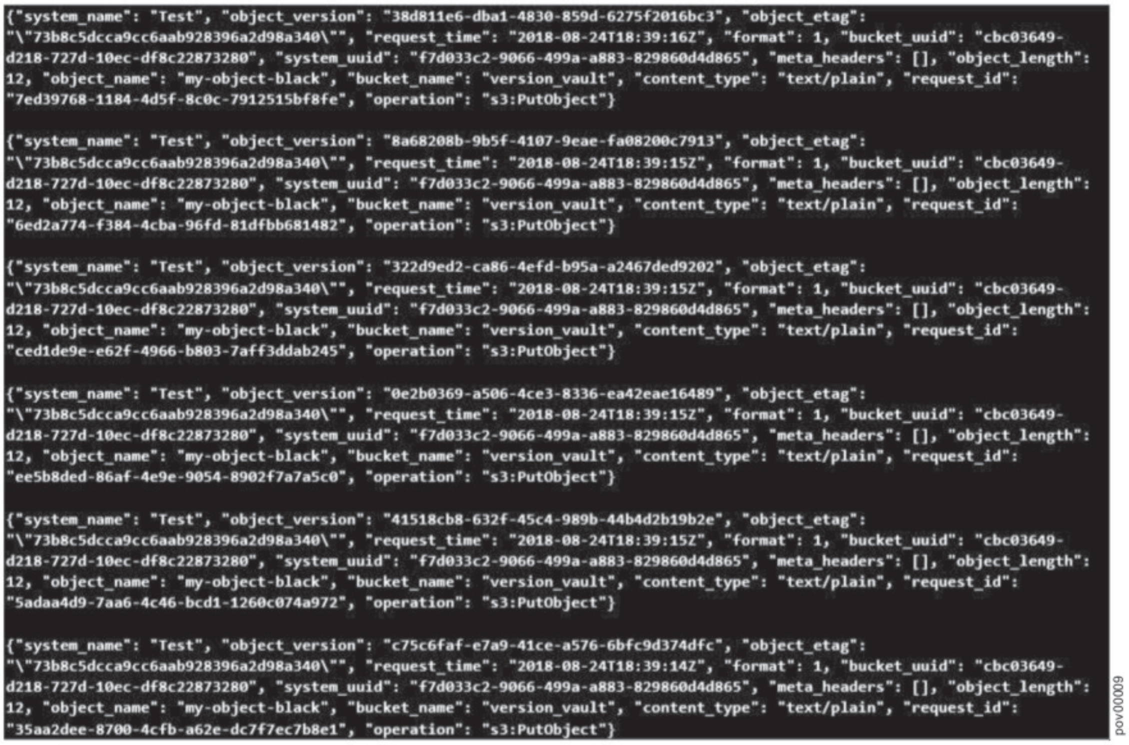
Table 25. Leaf directory file names (continued)

<table>
<tr><th>File name</th><th>Description</th></tr>
<tr><td>task.stats</td><td>Scanner starts in JSON format for a single vault. Updated following successful processing of each batch of objects.

"estimated_object_count": 1718,
"list_objects_size": 50,
"next_key": "",
"next_version": "",
"prefix": "",
"scan_type": "Object Scan",
"status": "Complete",
"total_bytes_output": 45257,
"total_bytes_scanned": 45257,
"total_objects_output": 1717,
"total_objects_scanned": 1717,
"vault_name": "test_vault2",
"vault_uuid": "06c1641d-082f-7ba2-011b-c7550651a780"</td></tr>
<tr><td>*.log</td><td>The Scanner creates multiple .log files for each vault. Each .log file contains up to 1000 Kafka messages, ready to be submitted to the Kafka cluster by the Notifier.
The naming convention for the log files is
<date>-<time>-<milliseconds>-<batch number>-
<number
of messages in file>.log

output
archive
data
Scenario2
Scenario3
Scenario4
Scenario5
Scenario6
Scenario7
test_vault2
v1
v2
version_vault
Name
_LISTProcess-3.debug
_task.stats
20180912-154817-798000-1-50.log
20180912-154818-160000-2-50.log
20180912-154818-468000-3-50.log
20180912-154818-775000-4-50.log
20180912-154819-112000-5-50.log</td></tr>
</table>

Appendix
The appendix shows an example of a log file and examples of Scanner debug data.

Log file

Figure 25 on page 123 shows an example with extra line breaks.

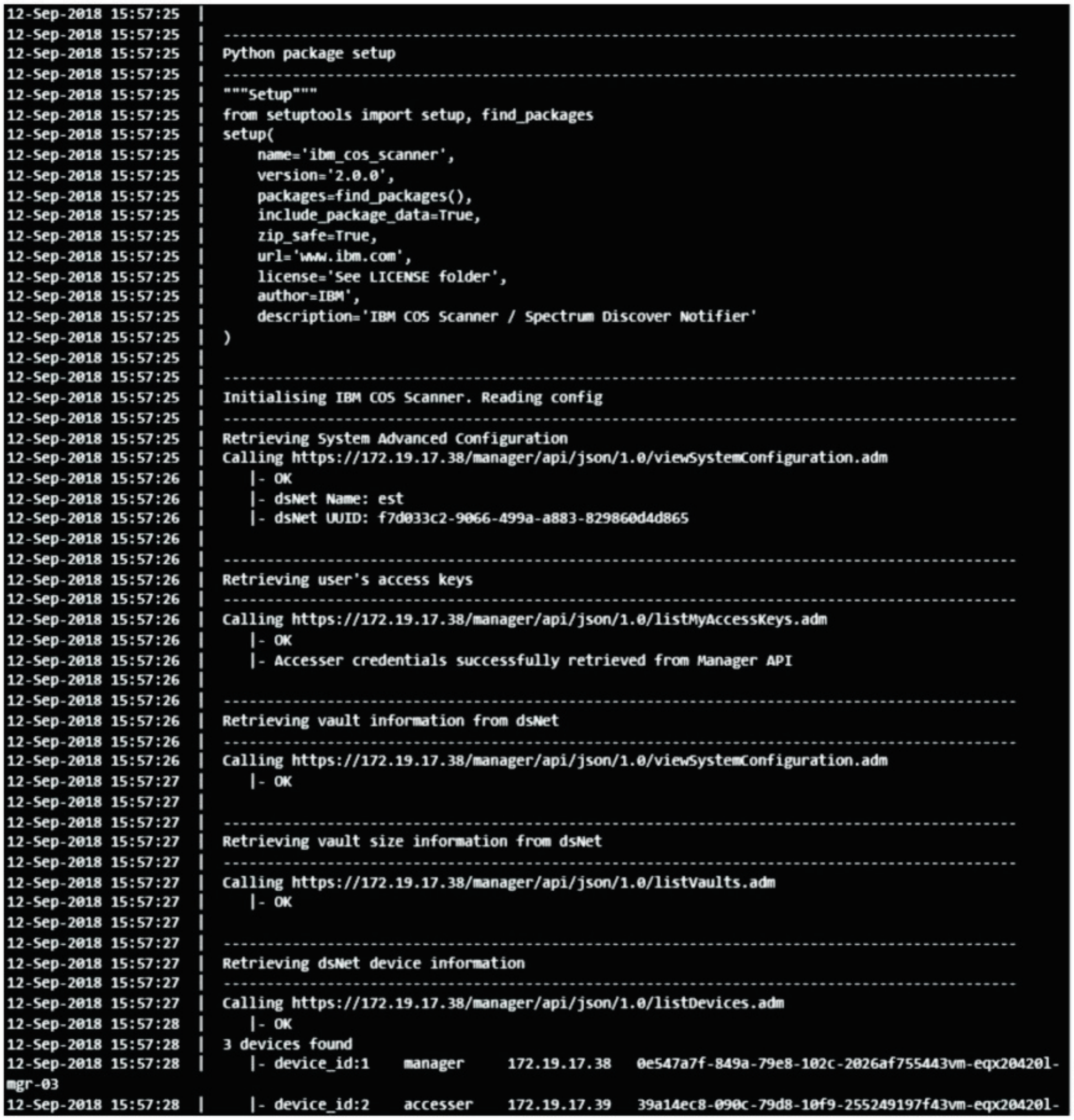


```
{"system_name": "Test", "object_version": "38d811e6-dba1-4830-859d-6275f2016bc3", "object_etag":
"\"73b8c5dcca9cc6aab928396a2d98a340\"", "request_time": "2018-08-24T18:39:16Z", "format": 1, "bucket_uuid": "cbc03649-
d218-727d-10ec-df8c22873280", "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865", "meta_headers": [], "object_length":
12, "object_name": "my-object-black", "bucket_name": "version_vault", "content_type": "text/plain", "request_id":
"7ed39768-1184-4d5f-8c0c-7912515bf8fe", "operation": "s3:PutObject"}

{"system_name": "Test", "object_version": "8a68208b-9b5f-4107-9eae-fa08200c7913", "object_etag":
"\"73b8c5dcca9cc6aab928396a2d98a340\"", "request_time": "2018-08-24T18:39:15Z", "format": 1, "bucket_uuid": "cbc03649-
d218-727d-10ec-df8c22873280", "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865", "meta_headers": [], "object_length":
12, "object_name": "my-object-black", "bucket_name": "version_vault", "content_type": "text/plain", "request_id":
"6ed2a774-f384-4cba-96fd-81dfbb681482", "operation": "s3:PutObject"}

{"system_name": "Test", "object_version": "322d9ed2-ca86-4efd-b95a-a2467ded9202", "object_etag":
"\"73b8c5dcca9cc6aab928396a2d98a340\"", "request_time": "2018-08-24T18:39:15Z", "format": 1, "bucket_uuid": "cbc03649-
d218-727d-10ec-df8c22873280", "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865", "meta_headers": [], "object_length":
12, "object_name": "my-object-black", "bucket_name": "version_vault", "content_type": "text/plain", "request_id":
"ced1de9e-e62f-4966-b803-7aff3ddab245", "operation": "s3:PutObject"}

{"system_name": "Test", "object_version": "0e2b0369-a506-4ce3-8336-ea42eae16489", "object_etag":
"\"73b8c5dcca9cc6aab928396a2d98a340\"", "request_time": "2018-08-24T18:39:15Z", "format": 1, "bucket_uuid": "cbc03649-
d218-727d-10ec-df8c22873280", "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865", "meta_headers": [], "object_length":
12, "object_name": "my-object-black", "bucket_name": "version_vault", "content_type": "text/plain", "request_id":
"ee5b8ded-86af-4e9e-9054-8902f7a7a5c0", "operation": "s3:PutObject"}

{"system_name": "Test", "object_version": "41518cb8-632f-45c4-989b-44b4d2b19b2e", "object_etag":
"\"73b8c5dcca9cc6aab928396a2d98a340\"", "request_time": "2018-08-24T18:39:15Z", "format": 1, "bucket_uuid": "cbc03649-
d218-727d-10ec-df8c22873280", "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865", "meta_headers": [], "object_length":
12, "object_name": "my-object-black", "bucket_name": "version_vault", "content_type": "text/plain", "request_id":
"5adaa4d9-7aa6-4c46-bcd1-1260c074a972", "operation": "s3:PutObject"}

{"system_name": "Test", "object_version": "c75c6faf-e7a9-41ce-a576-6bfc9d374dfc", "object_etag":
"\"73b8c5dcca9cc6aab928396a2d98a340\"", "request_time": "2018-08-24T18:39:14Z", "format": 1, "bucket_uuid": "cbc03649-
d218-727d-10ec-df8c22873280", "system_uuid": "f7d033c2-9066-499a-a883-829860d4d865", "meta_headers": [], "object_length":
12, "object_name": "my-object-black", "bucket_name": "version_vault", "content_type": "text/plain", "request_id":
"35aa2dee-8700-4cfb-a62e-dc7f7ec7b8e1", "operation": "s3:PutObject"}
```

Figure 25. Example of a log file

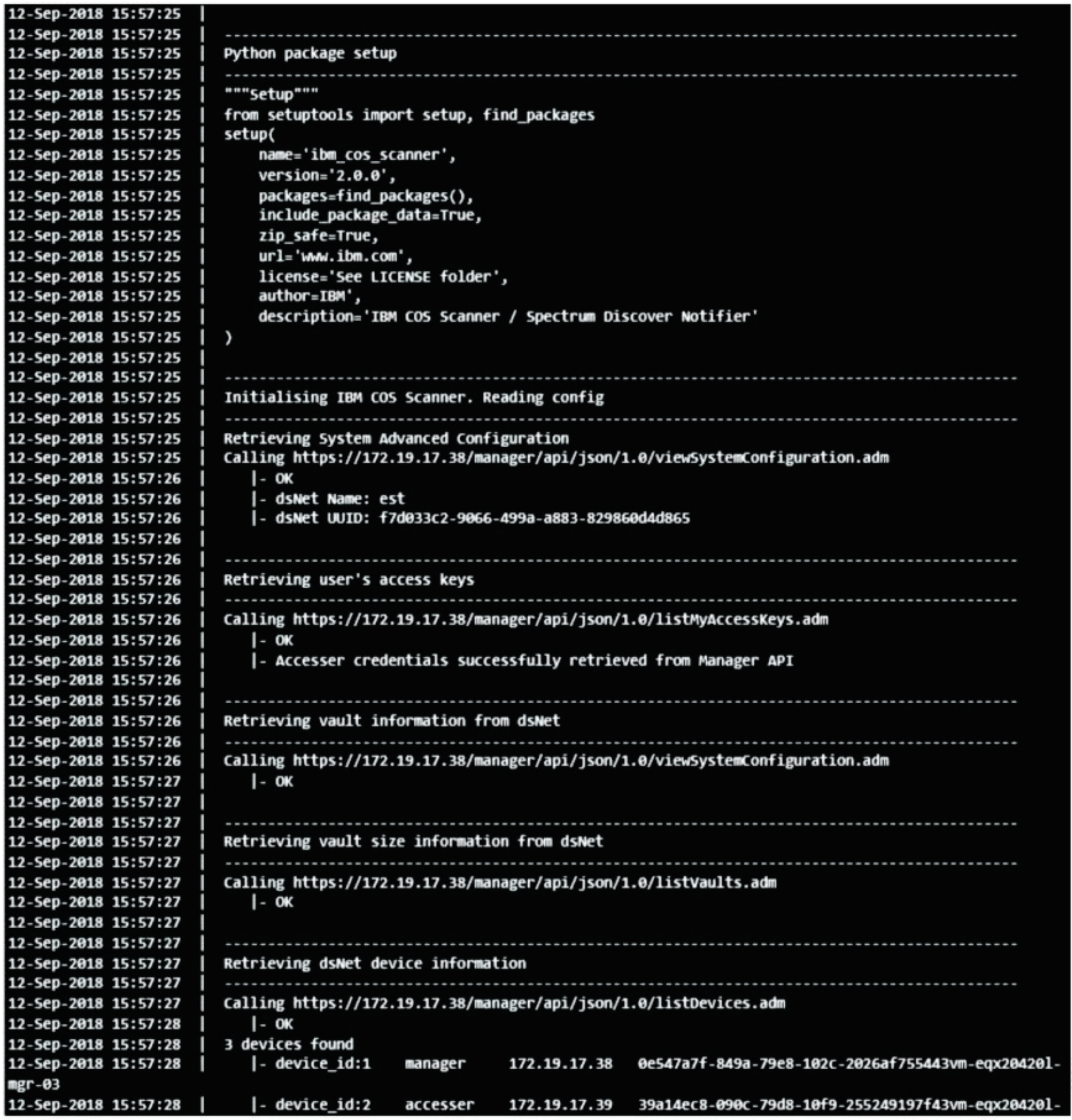
Scanner debug data

Figure 26 on page 124, Figure 27 on page 125, Figure 28 on page 126, and Figure 29 on page 127 show that throttling settings are logged and multiple HEAD processes are started for each LIST process.



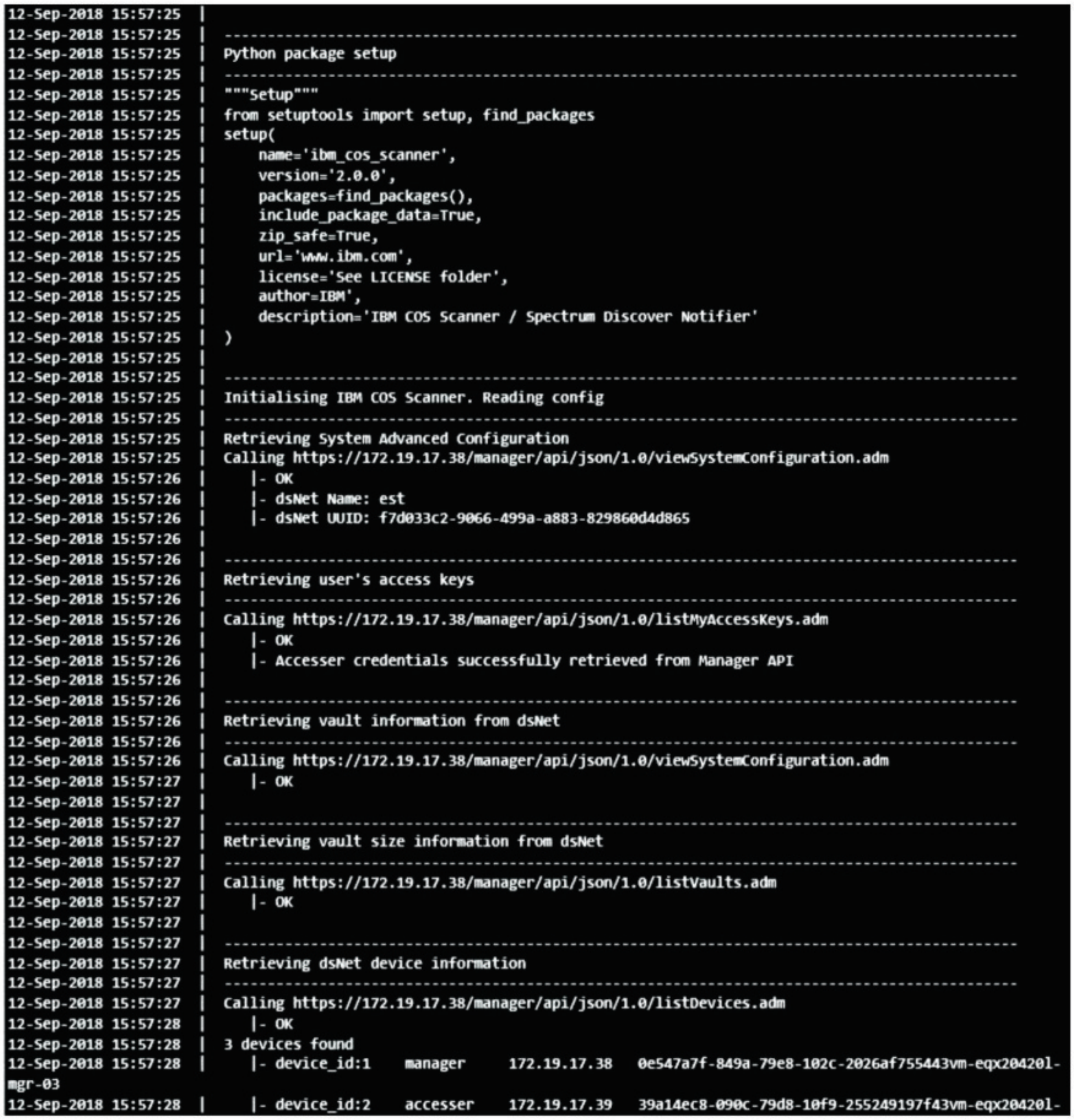
```
12-Sep-2018 15:57:25 |
12-Sep-2018 15:57:25 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Python package setup
12-Sep-2018 15:57:25 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  """Setup"""
12-Sep-2018 15:57:25 |  from setuptools import setup, find_packages
12-Sep-2018 15:57:25 |  setup(
12-Sep-2018 15:57:25 |      name='ibm_cos_scanner',
12-Sep-2018 15:57:25 |      version='2.0.0',
12-Sep-2018 15:57:25 |      packages=find_packages(),
12-Sep-2018 15:57:25 |      include_package_data=True,
12-Sep-2018 15:57:25 |      zip_safe=True,
12-Sep-2018 15:57:25 |      url='www.ibm.com',
12-Sep-2018 15:57:25 |      license='See LICENSE folder',
12-Sep-2018 15:57:25 |      author=IBM',
12-Sep-2018 15:57:25 |      description='IBM COS Scanner / Spectrum Discover Notifier'
12-Sep-2018 15:57:25 |  )
12-Sep-2018 15:57:25 |
12-Sep-2018 15:57:25 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Initialising IBM COS Scanner. Reading config
12-Sep-2018 15:57:25 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Retrieving System Advanced Configuration
12-Sep-2018 15:57:25 |  Calling https://172.19.17.38/manager/api/json/1.0/viewSystemConfiguration.adm
12-Sep-2018 15:57:26 |     |- OK
12-Sep-2018 15:57:26 |     |- dsNet Name: est
12-Sep-2018 15:57:26 |     |- dsNet UUID: f7d033c2-9066-499a-a883-829860d4d865
12-Sep-2018 15:57:26 |
12-Sep-2018 15:57:26 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Retrieving user's access keys
12-Sep-2018 15:57:26 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Calling https://172.19.17.38/manager/api/json/1.0/listMyAccessKeys.adm
12-Sep-2018 15:57:26 |     |- OK
12-Sep-2018 15:57:26 |     |- Accesser credentials successfully retrieved from Manager API
12-Sep-2018 15:57:26 |
12-Sep-2018 15:57:26 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Retrieving vault information from dsNet
12-Sep-2018 15:57:26 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Calling https://172.19.17.38/manager/api/json/1.0/viewSystemConfiguration.adm
12-Sep-2018 15:57:27 |     |- OK
12-Sep-2018 15:57:27 |
12-Sep-2018 15:57:27 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Retrieving vault size information from dsNet
12-Sep-2018 15:57:27 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Calling https://172.19.17.38/manager/api/json/1.0/listVaults.adm
12-Sep-2018 15:57:27 |     |- OK
12-Sep-2018 15:57:27 |
12-Sep-2018 15:57:27 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Retrieving dsNet device information
12-Sep-2018 15:57:27 |  ------------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Calling https://172.19.17.38/manager/api/json/1.0/listDevices.adm
12-Sep-2018 15:57:28 |     |- OK
12-Sep-2018 15:57:28 |  3 devices found
12-Sep-2018 15:57:28 |     |- device_id:1    manager     172.19.17.38   0e547a7f-849a-79e8-102c-2026af755443vm-eqx204201-
mgr-03
12-Sep-2018 15:57:28 |     |- device_id:2    accesser    172.19.17.39   39a14ec8-090c-79d8-10f9-255249197f43vm-eqx204201-
```

Figure 26. Scanner debug



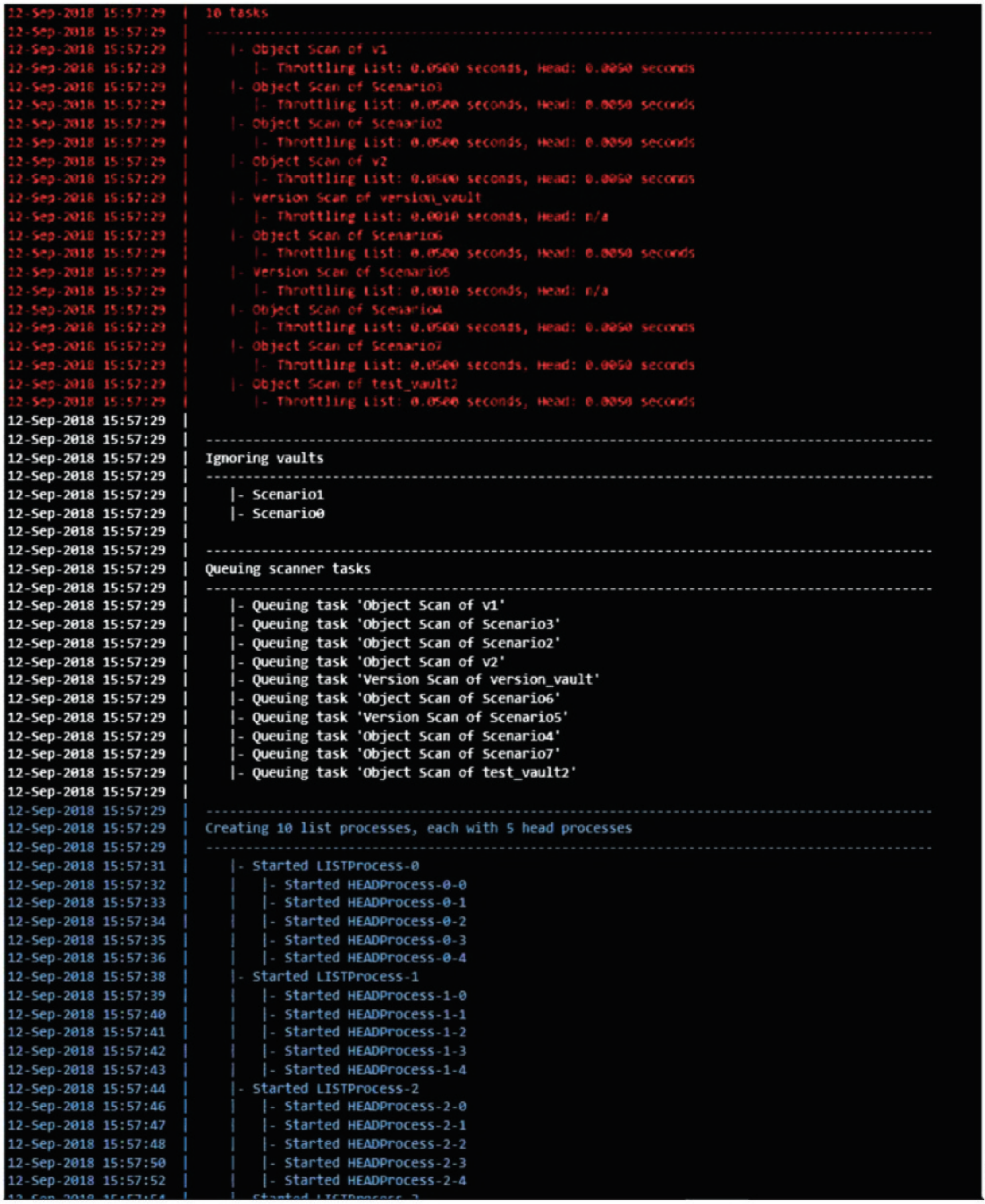
```
12-Sep-2018 15:57:25 |
12-Sep-2018 15:57:25 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Python package setup
12-Sep-2018 15:57:25 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  """Setup"""
12-Sep-2018 15:57:25 |  from setuptools import setup, find_packages
12-Sep-2018 15:57:25 |  setup(
12-Sep-2018 15:57:25 |      name='ibm_cos_scanner',
12-Sep-2018 15:57:25 |      version='2.0.0',
12-Sep-2018 15:57:25 |      packages=find_packages(),
12-Sep-2018 15:57:25 |      include_package_data=True,
12-Sep-2018 15:57:25 |      zip_safe=True,
12-Sep-2018 15:57:25 |      url='www.ibm.com',
12-Sep-2018 15:57:25 |      license='See LICENSE folder',
12-Sep-2018 15:57:25 |      author=IBM',
12-Sep-2018 15:57:25 |      description='IBM COS Scanner / Spectrum Discover Notifier'
12-Sep-2018 15:57:25 |  )
12-Sep-2018 15:57:25 |
12-Sep-2018 15:57:25 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Initialising IBM COS Scanner. Reading config
12-Sep-2018 15:57:25 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Retrieving System Advanced Configuration
12-Sep-2018 15:57:25 |  Calling https://172.19.17.38/manager/api/json/1.0/viewSystemConfiguration.adm
12-Sep-2018 15:57:26 |     |- OK
12-Sep-2018 15:57:26 |     |- dsNet Name: est
12-Sep-2018 15:57:26 |     |- dsNet UUID: f7d033c2-9066-499a-a883-829860d4d865
12-Sep-2018 15:57:26 |
12-Sep-2018 15:57:26 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Retrieving user's access keys
12-Sep-2018 15:57:26 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Calling https://172.19.17.38/manager/api/json/1.0/listMyAccessKeys.adm
12-Sep-2018 15:57:26 |     |- OK
12-Sep-2018 15:57:26 |     |- Accesser credentials successfully retrieved from Manager API
12-Sep-2018 15:57:26 |
12-Sep-2018 15:57:26 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Retrieving vault information from dsNet
12-Sep-2018 15:57:26 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Calling https://172.19.17.38/manager/api/json/1.0/viewSystemConfiguration.adm
12-Sep-2018 15:57:27 |     |- OK
12-Sep-2018 15:57:27 |
12-Sep-2018 15:57:27 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Retrieving vault size information from dsNet
12-Sep-2018 15:57:27 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Calling https://172.19.17.38/manager/api/json/1.0/listVaults.adm
12-Sep-2018 15:57:27 |     |- OK
12-Sep-2018 15:57:27 |
12-Sep-2018 15:57:27 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Retrieving dsNet device information
12-Sep-2018 15:57:27 |  --------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Calling https://172.19.17.38/manager/api/json/1.0/listDevices.adm
12-Sep-2018 15:57:28 |     |- OK
12-Sep-2018 15:57:28 |  3 devices found
12-Sep-2018 15:57:28 |     |- device_id:1    manager     172.19.17.38   0e547a7f-849a-79e8-102c-2026af755443vm-eqx20420l-
mgr-03
12-Sep-2018 15:57:28 |     |- device_id:2    accesser    172.19.17.39   39a14ec8-090c-79d8-10f9-255249197f43vm-eqx20420l-
```

Figure 27. Scanner debug (continued)



```
12-Sep-2018 15:57:25 |
12-Sep-2018 15:57:25 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Python package setup
12-Sep-2018 15:57:25 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  """Setup"""
12-Sep-2018 15:57:25 |  from setuptools import setup, find_packages
12-Sep-2018 15:57:25 |  setup(
12-Sep-2018 15:57:25 |      name='ibm_cos_scanner',
12-Sep-2018 15:57:25 |      version='2.0.0',
12-Sep-2018 15:57:25 |      packages=find_packages(),
12-Sep-2018 15:57:25 |      include_package_data=True,
12-Sep-2018 15:57:25 |      zip_safe=True,
12-Sep-2018 15:57:25 |      url='www.ibm.com',
12-Sep-2018 15:57:25 |      license='See LICENSE folder',
12-Sep-2018 15:57:25 |      author=IBM',
12-Sep-2018 15:57:25 |      description='IBM COS Scanner / Spectrum Discover Notifier'
12-Sep-2018 15:57:25 |  )
12-Sep-2018 15:57:25 |
12-Sep-2018 15:57:25 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Initialising IBM COS Scanner. Reading config
12-Sep-2018 15:57:25 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:25 |  Retrieving System Advanced Configuration
12-Sep-2018 15:57:25 |  Calling https://172.19.17.38/manager/api/json/1.0/viewSystemConfiguration.adm
12-Sep-2018 15:57:26 |     |- OK
12-Sep-2018 15:57:26 |     |- dsNet Name: est
12-Sep-2018 15:57:26 |     |- dsNet UUID: f7d033c2-9066-499a-a883-829860d4d865
12-Sep-2018 15:57:26 |
12-Sep-2018 15:57:26 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Retrieving user's access keys
12-Sep-2018 15:57:26 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Calling https://172.19.17.38/manager/api/json/1.0/listMyAccessKeys.adm
12-Sep-2018 15:57:26 |     |- OK
12-Sep-2018 15:57:26 |     |- Accesser credentials successfully retrieved from Manager API
12-Sep-2018 15:57:26 |
12-Sep-2018 15:57:26 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Retrieving vault information from dsNet
12-Sep-2018 15:57:26 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:26 |  Calling https://172.19.17.38/manager/api/json/1.0/viewSystemConfiguration.adm
12-Sep-2018 15:57:27 |     |- OK
12-Sep-2018 15:57:27 |
12-Sep-2018 15:57:27 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Retrieving vault size information from dsNet
12-Sep-2018 15:57:27 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Calling https://172.19.17.38/manager/api/json/1.0/listVaults.adm
12-Sep-2018 15:57:27 |     |- OK
12-Sep-2018 15:57:27 |
12-Sep-2018 15:57:27 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Retrieving dsNet device information
12-Sep-2018 15:57:27 |  -------------------------------------------------------------------------------------
12-Sep-2018 15:57:27 |  Calling https://172.19.17.38/manager/api/json/1.0/listDevices.adm
12-Sep-2018 15:57:28 |     |- OK
12-Sep-2018 15:57:28 |  3 devices found
12-Sep-2018 15:57:28 |     |- device_id:1    manager     172.19.17.38   0e547a7f-849a-79e8-102c-2026af755443vm-eqx20420l-
mgr-03
12-Sep-2018 15:57:28 |     |- device_id:2    accesser    172.19.17.39   39a14ec8-090c-79d8-10f9-255249197f43vm-eqx20420l-
```

Figure 28. Scanner debug (continued)



```
12-Sep-2018 15:57:29 |  10 tasks
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:29 |     |- Object Scan of v1
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Object Scan of Scenario3
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Object Scan of Scenario2
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Object Scan of v2
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Version Scan of version_vault
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0010 seconds, Head: n/a
12-Sep-2018 15:57:29 |     |- Object Scan of Scenario6
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Version Scan of Scenario5
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0010 seconds, Head: n/a
12-Sep-2018 15:57:29 |     |- Object Scan of Scenario4
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Object Scan of Scenario7
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds. Head: 0.0050 seconds
12-Sep-2018 15:57:29 |     |- Object Scan of test_vault2
12-Sep-2018 15:57:29 |        |- Throttling List: 0.0500 seconds, Head: 0.0050 seconds
12-Sep-2018 15:57:29 |
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:29 |  Ignoring vaults
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:29 |     |- Scenario1
12-Sep-2018 15:57:29 |     |- Scenario0
12-Sep-2018 15:57:29 |
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:29 |  Queuing scanner tasks
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of v1'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of Scenario3'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of Scenario2'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of v2'
12-Sep-2018 15:57:29 |     |- Queuing task 'Version Scan of version_vault'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of Scenario6'
12-Sep-2018 15:57:29 |     |- Queuing task 'Version Scan of Scenario5'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of Scenario4'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of Scenario7'
12-Sep-2018 15:57:29 |     |- Queuing task 'Object Scan of test_vault2'
12-Sep-2018 15:57:29 |
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:29 |  Creating 10 list processes, each with 5 head processes
12-Sep-2018 15:57:29 |  -----------------------------------------------------------------------------------
12-Sep-2018 15:57:31 |     |- Started LISTProcess-0
12-Sep-2018 15:57:32 |     |    |- Started HEADProcess-0-0
12-Sep-2018 15:57:33 |     |    |- Started HEADProcess-0-1
12-Sep-2018 15:57:34 |     |    |- Started HEADProcess-0-2
12-Sep-2018 15:57:35 |     |    |- Started HEADProcess-0-3
12-Sep-2018 15:57:36 |     |    |- Started HEADProcess-0-4
12-Sep-2018 15:57:38 |     |- Started LISTProcess-1
12-Sep-2018 15:57:39 |     |    |- Started HEADProcess-1-0
12-Sep-2018 15:57:40 |     |    |- Started HEADProcess-1-1
12-Sep-2018 15:57:41 |     |    |- Started HEADProcess-1-2
12-Sep-2018 15:57:42 |     |    |- Started HEADProcess-1-3
12-Sep-2018 15:57:43 |     |    |- Started HEADProcess-1-4
12-Sep-2018 15:57:44 |     |- Started LISTProcess-2
12-Sep-2018 15:57:46 |     |    |- Started HEADProcess-2-0
12-Sep-2018 15:57:47 |     |    |- Started HEADProcess-2-1
12-Sep-2018 15:57:48 |     |    |- Started HEADProcess-2-2
12-Sep-2018 15:57:50 |     |    |- Started HEADProcess-2-3
12-Sep-2018 15:57:52 |     |    |- Started HEADProcess-2-4
12-Sep-2018 15:57:54 |     |- Started LISTProcess-3
```

Figure 29. Scanner debug (continued)

Configure IBM Cloud Object Storage notifications for IBM Spectrum Discover

Ingesting IBM Cloud Object Storage event records into IBM Spectrum Discover requires the user to enable the Notification service on the IBM Cloud Object Storage system. Thereafter, the user must connect the IBM Cloud Object Storage system to the IBM Cloud Object Storage connector Kafka topic on the IBM Spectrum Discover cluster. The name of this connector topic is `cos-le-connector-topic`.

A combination of SASL and TLS is used to authenticate and encrypt the connection between the IBM Cloud Object Storage source system and the Kafka brokers which reside in the IBM Spectrum Discover cluster. The certificate and credentials required to establish this connection might be obtained directly from the IBM Spectrum Discover cluster by the IBM Spectrum Discover storage administrator.

For information on how to enable and configure the IBM Cloud Object Storage Notification service with the IBM Spectrum Discover provided credentials, see IBM Cloud Object Storage Administration Documentation.

The following information is required to establish a secure connection between IBM Cloud Object Storage and IBM Spectrum Discover:

Hosts
: One or more of the IBM Spectrum Discover Kafka brokers is in the format: *host1:port,host2:port*. The Kafka producers on the IBM Cloud Object Storage system will retrieve the full list of IBM Spectrum Discover Kafka brokers from the first host that is alive and responding. The broker's host and port (the list configured might contain more than one broker) for SASL SSL can be obtained by the IBM Spectrum Discover storage administrator from the following location on the IBM Spectrum Discover master node: `/etc/kafka/server.properties`.

Authentication credentials
: The user name is `cos` and the password can be obtained by the IBM Spectrum Discover storage administrator from the following location on the IBM Spectrum Discover master node: `/etc/kafka/sasl_password`.

Certificate PEM for TLS encryption
: This the CA certificate that is used to sign the Kafka server and client certificates for the IBM Spectrum Discover cluster. It might be obtained by the IBM Spectrum Discover storage administrator from the following location on the IBM Spectrum Discover master node: `/etc/kafka/ca.crt`.

  This file is in the PEM format and the entirety of its contents must be pasted into the **Certificate PEM** field of the **COS Notifications** configuration panel.

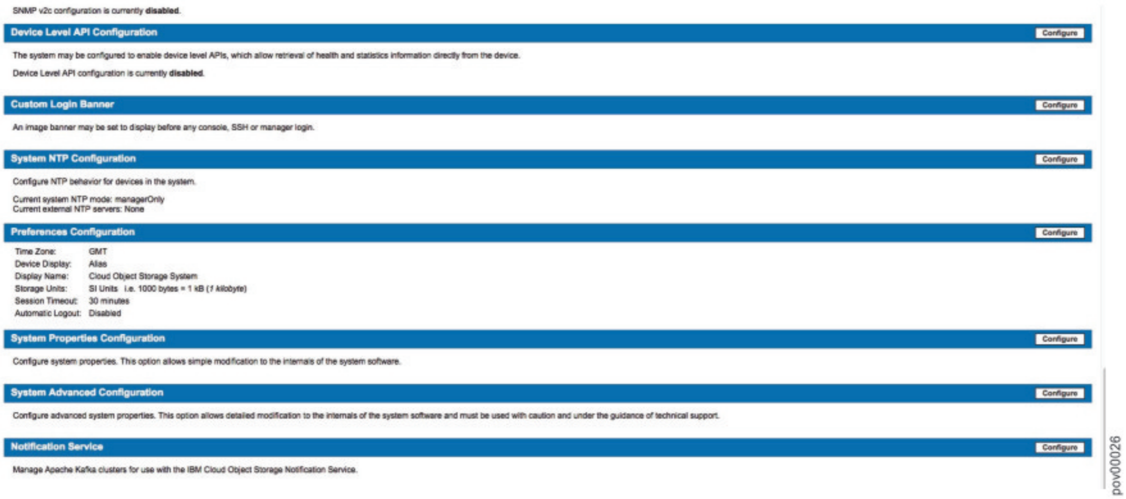
Enabling IBM Cloud Object Storage notification services

The IBM Cloud Object Storage notification service can be enabled with the information that follows:

Procedure

1. Log in to the IBM Cloud Object Storage Manager Admin console https://manager_host/manager/login.adm with a user name of **`admin`** and a password of **`password`**.

   If you defined your own password, use your pre-defined password. If you do not have a pre-defined password, use the default password.
2. Select the **Administration** tab.
3. Scroll to the end of the page and select **Configure the Notification Service**.



SNMP v2c configuration is currently **disabled**.

Device Level API Configuration

The system may be configured to enable device level APIs, which allow retrieval of health and statistics information directly from the device.

Device Level API configuration is currently **disabled**.

Custom Login Banner

An image banner may be set to display before any console, SSH or manager login.

System NTP Configuration

Configure NTP behavior for devices in the system.

Current system NTP mode: managerOnly
Current external NTP servers: None

Preferences Configuration

Time Zone: GMT
Device Display: Alias
Display Name: Cloud Object Storage System
Storage Units: SI Units i.e. 1000 bytes = 1 kB (*1 kilobyte*)
Session Timeout: 30 minutes
Automatic Logout: Disabled

System Properties Configuration

Configure system properties. This option allows simple modification to the internals of the system software.

System Advanced Configuration

Configure advanced system properties. This option allows detailed modification to the internals of the system software and must be used with caution and under the guidance of technical support.

Notification Service

Manage Apache Kafka clusters for use with the IBM Cloud Object Storage Notification Service.

Figure 30. Configurations

   Before you add a notification service to the IBM Cloud Object Storage platform, you must obtain some information from the IBM Spectrum Discover server.

To authenticate the IBM Cloud Object Storage notification service, you can capture the Kafka user name and password from the files on the IBM Spectrum Discover platform.

If the Transport Layer Security (TLS) is enabled in the IBM Cloud Object Storage notification service, you can also copy the certificate authority (CA) in PEM format from the IBM Spectrum Discover platform. After you collect the information, you can add the information to the notification service configuration.

Authenticating, encrypting, and enabling

Log in to the IBM Spectrum Discover server to extract the information that follows:

1. Log in to the IBM Spectrum Discover server and extract the information from the screen below.
2. See the screen below for an example of Kafka user name and password.

```
moadmin@server kafka]$ cd /etc/kafka

[moadmin@server kafka]$ cat kafka_server_jaas.conf
KafkaServer {
org.apache.kafka.common.security.plain.PlainLoginModule required
user_cos="meezDMxFNZJMSxdyWQKSjVbs";
};

User= cos
Password = meezDMxFNZJMSxdyWQKSjVbs
```

Encryption

This topic shows an example of a certificate of authority for the PEM file.

1. Copy the block of text in the screen below starting with **BEGIN CERTIFICATE** and ending with **END CERTIFICATE**.

```
-----BEGIN CERTIFICATE-----
MIIExTCCA62gAwIBAgIJAKMX/n6ULb6YMA0GCSqGSIb3DQEBCwUAMIGYMQswCQYD
VQQGEwJHQjEOMAwGA1UECAwFSEFOVFMxEDAOBgNVBAcMB0h1cnNsZXkxDDAKBgNV
BAoMA0lCTTEZMBcGA1UECwwQc3BlY3RydW1kaXNjb3ZlcjEZMBcGA1UEAwwQc3Bl
Y3RydW1kaXNjb3ZlcjEjMCEGCSqGSIb3DQEJARYUbWxhd3JlbmNlQHVrLmlibS5j
b20wHhcNMTkwMTAyMTY1MDU5WhcNMzgxMjI4MTY1MDU5WjCBmDELMAkGA1UEBhMC
R0IxDjAMBgNVBAgMBUhBTlRTMRAwDgYDVQQHDAdIdXJzbGV5MQwwCgYDVQQKDANJ
Qk0xGTAXBgNVBAsMEHNwZWN0cnVtZGlzY292ZXIxGTAXBgNVBAMMEHNwZWN0cnVt
ZGlzY292ZXIxIzAhBgkqhkiG9w0BCQEWFG1sYXdyZW5jZUB1ay5pYm0uY29tMIIB
IjANBgkqhkiG9w0BAQEFAAOCAQ8AMIIBCgKCAQEAwg7z4gDeWlkeJjPvj3wobDBB
JrHJngooDbPLicRSf/yjl1NgwbWbjIjIeL9R8My+24hRUGfym9IWCM8qMWyEHG+w
+Rr/6jdQyD89j+m1c2ly3nDhXYsTQZR03UylC/TimF6fc07CfuQ1E2ljHf/JXVK4
ESVilhZR23/tWIfbITZmLvdftJSx0Kgu0Ow4BIr9kpQ3bXwt/eoDvAhdKztDouWN
lYCGmdzFOi6E3asspxHhcsGW3bcMu5mqzT6BEnSzrxr8kRbRDL6Q0Pqv33XVxP6z
OHIvv1uFg9Vq6XHIZLBhWNDqPgYoAbT0Q43vUxk7mJ3uJQY6bgbfuEa+PxygQwID
AQABo4IBDjCCAQowHQYDVR0OBBYEFEKxmmHeSfxgHuFL1dd82WMyf190MIHNBgNV
HSMEgcUwgcKAFEKxmmHeSfxgHuFL1dd82WMyf190oYGepIGbMIGYMQswCQYDVQQG
EwJHQjEOMAwGA1UECAwFSEFOVFMxEDAOBgNVBAcMB0h1cnNsZXkxDDAKBgNVBAoM
A0lCTTEZMBcGA1UECwwQc3BlY3RydW1kaXNjb3ZlcjEZMBcGA1UEAwwQc3BlY3Ry
dW1kaXNjb3ZlcjEjMCEGCSqGSIb3DQEJARYUbWxhd3JlbmNlQHVrLmlibS5jb22C
CQCjF/5+lC2+mDAMBgNVHRMEBTADAQH/MAsGA1UdDwQEAwIBBjANBgkqhkiG9w0B
AQsFAAOCAQEANINRvyeuJh69iRK5dPJssmcISXcZv4X33ukAyRt4zLNFToSkTfj2
ZAtQCNgQNl9Ln7Twuit+e6wifxAkA+UD7wrxMzb32+Mpw/XNzo5DnhInfvkAfC62
SHqWIaqTLXDeGbE8O7ieFsI7kAgEQCf23z/vESB2+m1XBI1UcuxMioYwX4YTb14/
GLDJkqhXMLWV+h/7NU7KbERSBia24N5zlR6Ed/rx83uD2AwBnBqt24sD6Q8Gbm+e
HLMv0JrH1vty1vGsfkZnSHb+E6V/5+GsnpIaDyIpsCvM1LqS/wMzBg9hlT5sii8l
mmqMTK6yqcqS7CfWFv/DjQr/i9ECyJ8fAQ==
-----END CERTIFICATE-----
```

Notification service configuration setup

1. Check **Enable Configuration**.

```
NAME: <NAME>
Topic: cos-le-connector-topic
Hosts: <SD ipaddress> :9092
Type: IBM Spectrum Discover
```

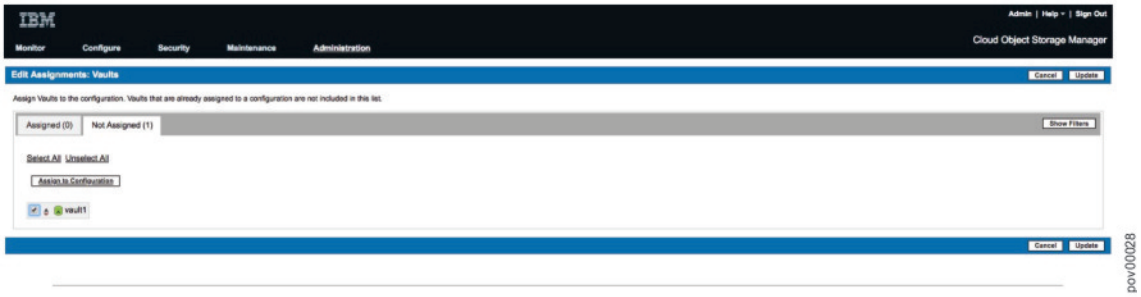
Enabling authentication

1. Check **Enable authentication**.

```
Username: cos
Password: <PASSWORD>
```

Enabling encryption

1. Check **Enable TLS for Apache Kafka network connections**.
2. Add the certificate PEM file from the IBM Spectrum Discover platform. See Figure 31 on page 130.



Assign Vaults to the configuration. Vaults that are already assigned to a configuration are not included in this list.

Figure 31. Add a storage vault to the configuration

Testing the IBM Cloud Object Storage notification service

To test the IBM Cloud Object Storage notification service, the tester can populate the IBM Cloud Object Storage vault with test data.

About this task

You can use a number of methods to write files to an IBM Cloud Object Storage vault, but you can use cURL directly on IBM Spectrum Discover platform. cURL is a computer software project that provides a library and command-line tool for transferring data that uses various protocols.

Procedure

1. Create a test file, for example, object_1.txt.

   The test file can be any file that contains data.
2. Write a file to the IBM Cloud Object Storage vault by using cURL.

   Requirements

   IBM COS Vault Name (vault1) [anonymous access enabled]
   IBM COS Accesser IP address

Example

```
[moadmin@spectrum_discover~]$ curl -X PUT -i -T object_1.txt http://9.11.200.208/vault1/
object_1.txt
HTTP/1.1 100 ContinueHTTP/1.1 200 OK
Date: Fri, 04 Jan 2019 13:21:14 Greenwich mean time
X-Clv-Request-Id: a9ad657a-a919-4b13-9b72-961ae8c57e3c
Server: 3.14.0.23
X-Clv-S3-Version: 2.5
x-amz-request-id: a9ad657a-a919-4b13-9b72-961ae8c57e3c
ETag: "7c517c7108f7180377e7b37db2e39261"
Content-Length: 0
```

Monitoring the IBM Cloud Object Storage accesser logs

To determine whether a file has been successfully written to the IBM Cloud Object Storage vault and a notification has been successfully sent to the IBM Spectrum Discover server, the accesser logs can be monitored on the IBM Cloud Object Storage Accesser server.

In the following example, an object that is written to vault1 results in the sending of one notification to the IBM Spectrum Discover server. The user must have access privileges to log on to the IBM Cloud Object Storage Accesser host to check the log files.

Confirm that an object is stored in the IBM Cloud Object Storage vault.

```
root@ibm_accesser:/var/log/dsnet-core# tail -f http.log
9.11.201.78 - ""  -  [04/Jan/2019:13:21:14 +0000] "PUT /vault1/object_1.txt HTTP/1.1" 200 0 "-"
"curl/7.29.0" 22
```

Confirm that a notification is sent to the IBM Spectrum Discover server.

```
root@ibm_accesser:/var/log/dsnet-core# tail -f notification.log
{"time":"2019-01-04T13:21:14.668Z","request_id":"a9ad657a-
a919-4b13-9b72-961ae8c57e3c","retried":true,"success":true,
"request_time":"2019-01-04T13:21:14.567Z","kafka_config_uuid":"d842c7a0-9c36-412e-8908-8ad5120a261
e","topic":"cos-le-connector-topic”
```

Monitoring the IBM Spectrum Discover producer IBM Cloud Object Storage logs

When the IBM Spectrum Discover server receives a notification from the IBM Cloud Object Storage platform, the IBM Spectrum Discover producer IBM Cloud Object Storage will record a transaction.

A successful notification is recorded as an offset value of one, when a notification is received from IBM Cloud Object Storage platform.

```
[moadmin@spectrum_discover]$ kubectl logs -f -n producercos  kindled-alligator-producer-cos-
producer-9f6966b4-8jsg7
break time. waiting for work...
2019-01-04 13:21:19.187 > offset_commit_cb: success, offsets:[{part: 0, offset: 1, err: none}]
```

Monitoring the IBM Spectrum Discover dashboard for IBM Cloud Object Storage ingestion

After IBM Cloud Object Storage notifications have been ingested from the IBM Cloud Object Storage platform, the IBM Spectrum Discover dashboard should display the total number of indexed records.

Note that the IBM Spectrum Discover dashboard can take approximately 30 minutes to display the total number of indexed records. See Figure 32 on page 132 .

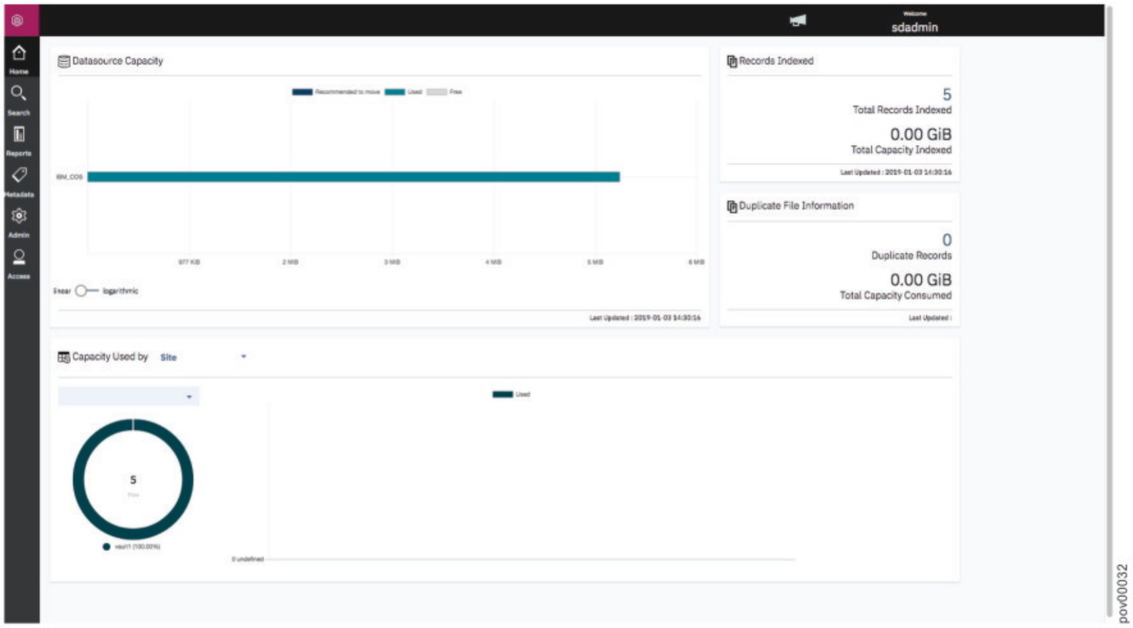


Figure 32. Total number of indexed records

IBM Spectrum Discover and S3 object storage data source connections

Use this information to understand how IBM Spectrum Discover works with S3-compliant object store.

Creating an S3 object storage data connection

Use this information to create a connection to an S3-compliant object store.

About this task

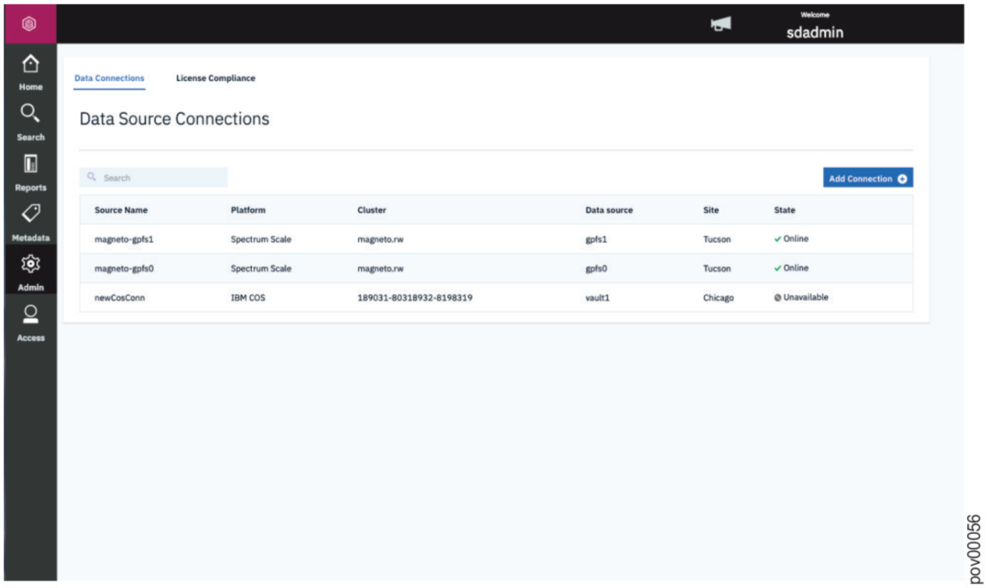
To create the S3-compliant connection:

Procedure

1. Log in to the IBM Spectrum Discover graphical user interface (GUI) with a user ID that is associated with data administration role.

   The data administration access role is required for creating connections. For more information about role-based access control, go to "Role-based access control" on page 3.
2. Select **Admin** from the left navigation menu.

   Clicking **Admin** displays the different types of data source connection names, platforms, clusters, data source, size, and **Add Connection**.



Data Source Connections

Source Name	Platform	Cluster	Data source	Site	State
magneto-gpfs1	Spectrum Scale	magneto.rw	gpfs1	Tucson	✓ Online
magneto-gpfs0	Spectrum Scale	magneto.rw	gpfs0	Tucson	✓ Online
newCosConn	IBM COS	189031-80318932-8198319	vault1	Chicago	● Unavailable

*Figure 33. Displaying the source names for **Data Source Connections***

3. Click **Add Connection** to display a new window that shows **Add Data Source Connection**.

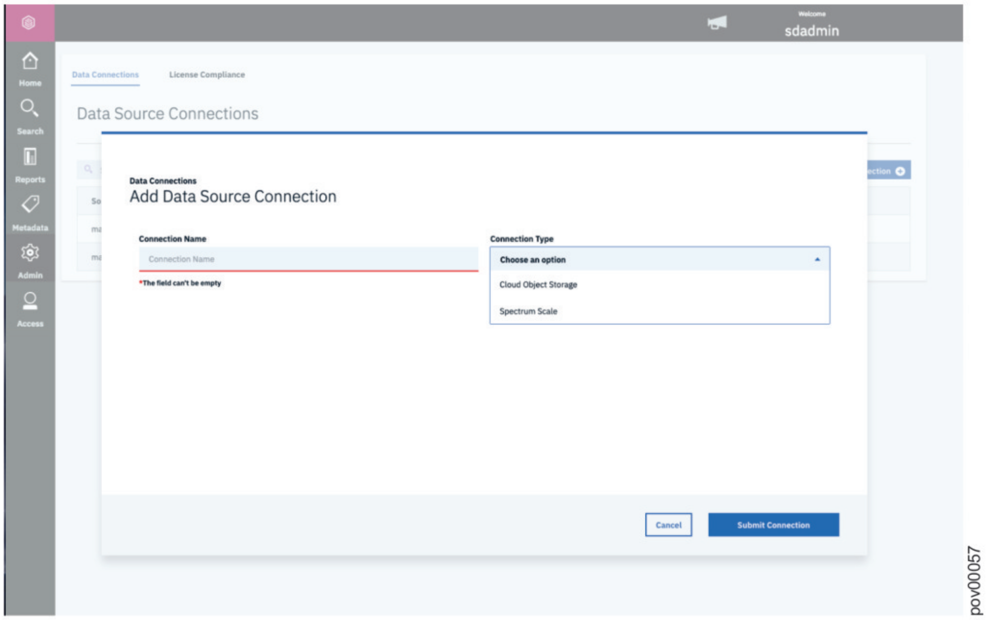


*Figure 34. Example of the **Add Data Source Connection** GUI window*

4. Complete the following steps:
   a) In the field for **Connection Name**, define a **Connection Name**.
   b) Click the down arrow for **Connection Type** to display a drop-down menu for **Choose an option**.
5. Select the connection type **S3 Object Storage**.

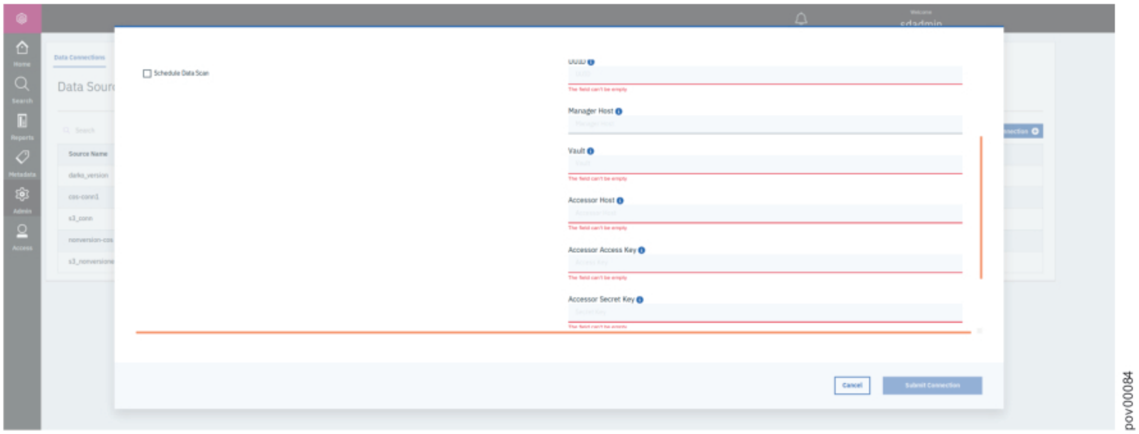


*Figure 35. Selecting **S3 Object Storage***

6. In the screen for S3, complete the fields and then click **Submit Connection** for the S3 connections manager.

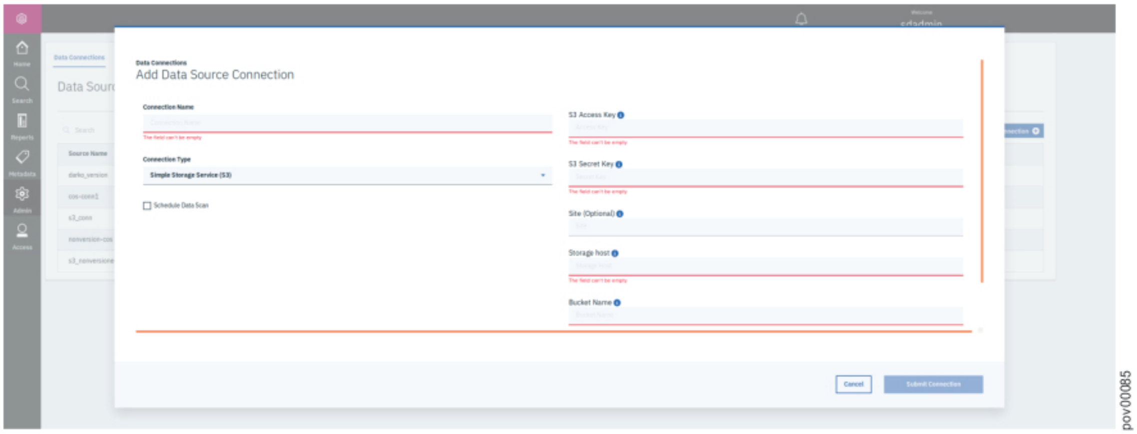


Figure 36. Completing the S3 information fields

S3 access key
: This is the access key ID for the S3 object store.

S3 secret key
: This is the secret key ID for the S3 object store.

Site
: This is the physical location of the records.

Storage host
: This is the IP address or host of the storage system.

Bucket name
: This is the name of the bucket that you are going to scan.

]

[Scanning an S3 object storage data connection

Use this information to scan a connection for an S3-compliant object store.

About this task

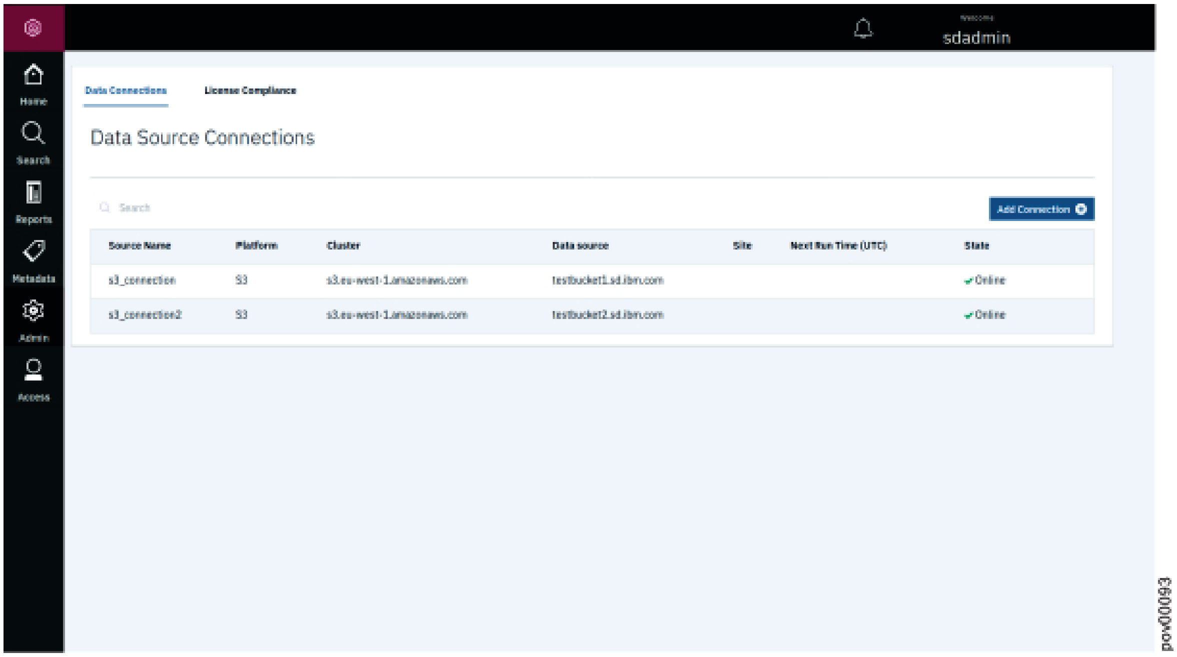
When you initiate a scan from the IBM Spectrum Discover graphical user interface (GUI), the metadata is transferred asynchronously back to the IBM Spectrum Discover instance.

Automated scanning and data ingestion relies on an established and active network connection between the IBM Spectrum Discover instance and the S3 storage source. If the connection cannot be established, the state of the data source connection shows as unavailable, and the option for automated scanning does not appear in the IBM Spectrum Discover GUI for that connection.

Procedure

1. Go to the IBM Spectrum Discover graphical user interface (GUI).
2. Under **Admin** select **Data Source connections**

   The following example shows the **Admin** data connections menu page:

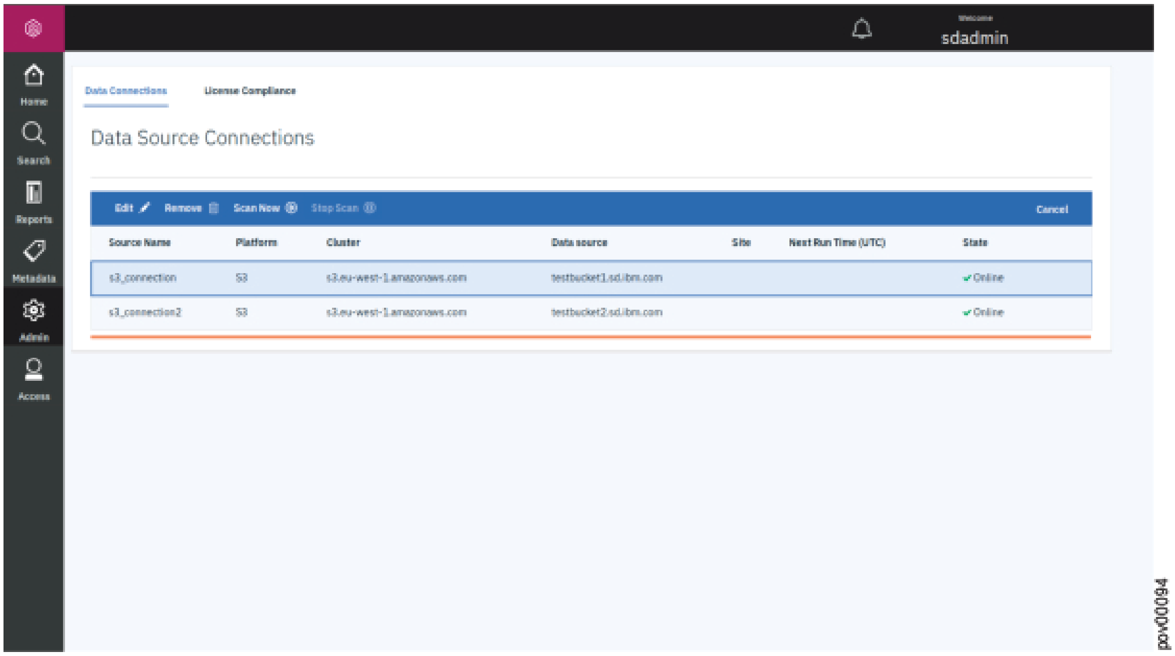
   

   Data Source Connections

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
s3_connection	S3	s3.eu-west-1.amazonaws.com	testbucket1.sd.ibm.com			✓Online
s3_connection2	S3	s3.eu-west-1.amazonaws.com	testbucket2.sd.ibm.com			✓Online

   Figure 37. Data source connections
3. Select the data source connection that you want to scan. Make sure that the `State` is listed as `Online` to make your system scan ready.

   The following example shows how to select a data source connection to scan.



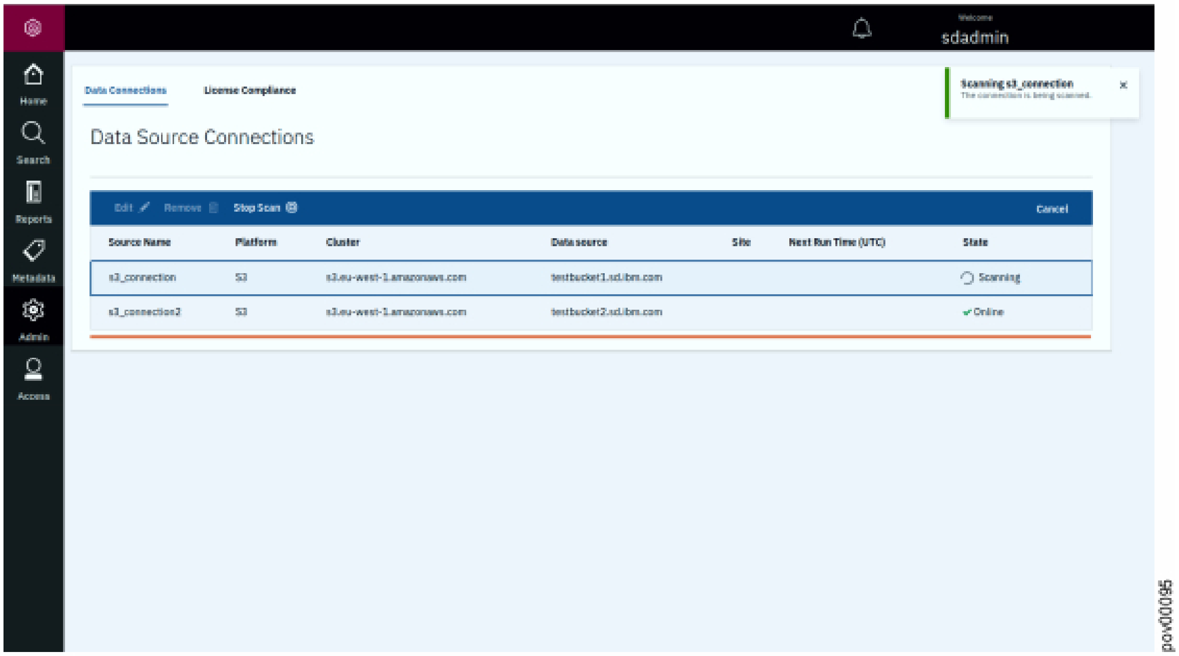
Data Source Connections

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
s3_connection	S3	s3.eu-west-1.amazonaws.com	testbucket1.sd.ibm.com			✓ Online
s3_connection2	S3	s3.eu-west-1.amazonaws.com	testbucket2.sd.ibm.com			✓ Online

Figure 38. Selecting a data source connection to scan

4. Select **Scan Now** to change the status to **Scanning**.

   The following example shows an active scan.



Scanning s3_connection
The connection is being scanned.

Data Source Connections

Source Name	Platform	Cluster	Data source	Site	Next Run Time (UTC)	State
s3_connection	S3	s3.eu-west-1.amazonaws.com	testbucket1.sd.ibm.com			Scanning
s3_connection2	S3	s3.eu-west-1.amazonaws.com	testbucket2.sd.ibm.com			✓ Online

Figure 39. Active scans

5. When the scan finishes, the state field returns to a status of **Online**.

Best practices for scanning an S3 object storage system

Use best practices for scanning S3-compliant object storage systems.

It is recommended to check the log files in the following directories after each scan:

`/gpfs/gpfs0/connections/s3/<connection_name>/debug/<scan_timestamp>/scanner.debug` indicates whether the scan was successful or not.

`/gpfs/gpfs0/connections/s3/<connection_name>/error/<scan_timestamp>/scanner.error` contains a list of all the messages that are not delivered to IBM Spectrum Discover

`/gpfs/gpfs0/connections/s3/<connection_name>/data/<scan_timestamp>/` contains a subfolder with the scanned data source name. There is a stats folder inside this folder that contains information about the number of objects in the data source or the number of objects or scanned files.

You can also compare the total size of the bucket that is reported in IBM Spectrum Discover with the total size of the S3 object store at source (if it is available).

]

Creating an NFS data source connection

Creating a Network File System (NFS) data connection using the IBM Spectrum Discover graphical user interface.

About this task

Note: For NFS scanning using Data ONTAP version 8.1.2 or later, the file system should be exported with the following configuration:

- Protocol - NFSv3
- Security flavor - UNIX
- Client permissions - A minimum of Read Only access is required
- Anonymous users - Root access must be granted
- Setuid and Setgid executable routines are not required

Creating data source connections in IBM Spectrum Discover identifies source storage systems that are to be indexed by IBM Spectrum Discover.

For some data source types, a network connection is (optionally) created to allow for automated scanning and indexing of the source system metadata. IBM Spectrum Discover will not index data from unknown sources, so creating a data source connection is the first step towards cataloging any source storage system.

Procedure

1. Log in to the IBM Spectrum Discover web interface with a user id that has the data admin role associated with it.

   Note: The data admin access role is required for creating connections.
2. Select **Admin** from the left navigation menu to display the source name, platform, cluster, data source, site, next run time, and state of existing data source connections.
3. Click **Add Connection** to display the **Add Data Source Connection** panel.
4. Enter a name in the **Connection Name** field.
5. Click **Connection Type** to expand the **Choose an option** menu.
6. Select the connection type **Network File System**, and click **Submit Connection**.
7. Fill in the fields for the Network File System parameters, and click **Submit Connection**.

Parameters for Network File System connections

Connection Name

The name of the connection, an identifier for the user, for example `filesystem1`.

Note: It must be a unique name within IBM Spectrum Discover.

Connection Type
The type of source storage system this connection represents.

Data Source
The full name of the data source.

Export Path
The data path from which data is to be exported.

Host
The host name of the node from which a scan can be initiated.

Site (Optional)
The city in which the data source facility is located.

Editing and using the TimeSinceAccess and SizeRange buckets

Users can group or aggregate data into two user-defined bucket ranges. The two user-defined bucket ranges are TimeSinceAccess and SizeRange. The TimeSinceAccess bucket groups files and objects based on the time they were last accessed.

The SizeRange bucket groups files and objects based on their size. Each of these buckets can be customized to better align with the user's requirements. Each bucket has up to five custom ranges with user-defined labels. [To access the SizeRange bucket groups, select **Metadata** > **Tags** > **edit icon for the SizeRange tag**.] For example, SizeRange can be broken up into 'T-shirt size' ranges where the ranges and labels are:

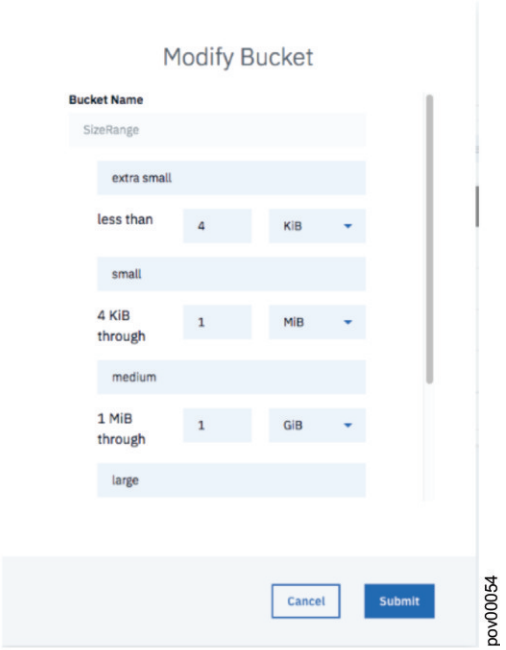
Table 26. Examples of size ranges and sizes of buckets with user-defined labels

Size range	Size
0 - 4 K	XS
4 K - 1 M	S
1 M - 1 G	M
1 G - 1T B	L
1 TB+	XL

[After you change or update a bucket definition, IBM Spectrum Discover summarizes the current set of files and objects into their respective bucket ranges. The changes are updated periodically every half an hour, thus it may take a half an hour or more before the changes are reflected in the Spectrum Discover GUI.]

Note: Ensure that the maximum value for each bucket is greater than the value assigned to the previous bucket.

See the Figure 40 on page 139



Modify Bucket

Bucket Name

SizeRange

extra small

less than 4 KiB

small

4 KiB through 1 MiB

medium

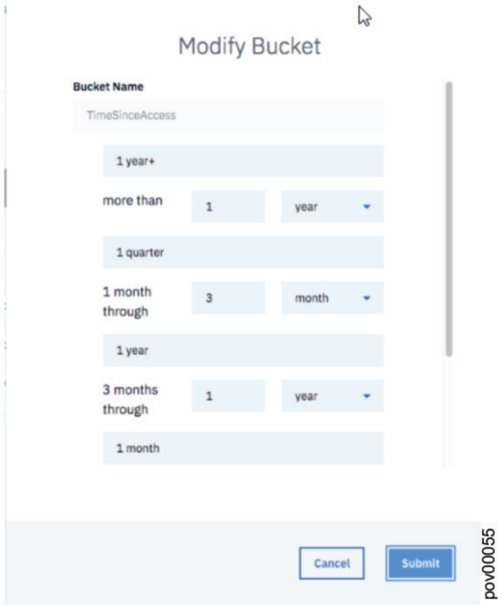
1 MiB through 1 GiB

large

Figure 40. Example of how to define the settings for a SizeRange bucket

[To open the menu for the TimeSinceAccess buckets select **Metadata** > **Tags** > **edit icon for the TimeSinceAccess tag**. See the Figure 40 on page 139 for an example.

Figure 41 on page 139 shows an example of how to modify and define the settings of a bucket that is older than one year.



Modify Bucket

Bucket Name

TimeSinceAccess

1 year+

more than 1 year

1 quarter

1 month through 3 month

1 year

3 months through 1 year

1 month

Figure 41. Example of how to modify and define the settings of a bucket that is older than one year old

Backup and restore

IBM Spectrum Discover includes a set of scripts for safely backing up and restoring your database and file system.

The scripts used to backup and restore databases and file systems are located in the **`/opt/ibm/metaocean/backup-restore`** directory, and must be run as root user (Example: **`sudo python /opt/ibm/metaocean/backup-restore/backup.py`**).

It is a good practice to back up your system at least once a week. IBM Spectrum Discover provides the **`automatedBackup.py`** script that can be used to configure a **`cron`** job that backs up your system and offloads a **`tar`** file to your selected storage server. The default configuration is daily at 12:00AM, however you can configure the backup frequency by running the **`automatedBackup.py`** script following the initial setup.

Remember: [

- If any files or a database become corrupted, run the **`restore.py`** script to recover your file system and database back to the date of your last successful backup.
- When you start a back or restore operation, remember that it can take up to 1 hour or more time to complete. Make sure that you plan accordingly.

]

[For more information, see the *IBM Spectrum Discover: Administration Guide*.]

Upgrading the IBM Spectrum Discover code

[You can upgrade the IBM Spectrum Discover code after you extract the tarball or tarfile to the `/opt/ibm/metaocean` directory on the master node.]

Procedure

1. [

   The IBM Spectrum Discover upgrade tool is a tarball or tarfile, which is the name of a group of files that are bundled together by using the tar command. You can download this tarball or tarfile from IBM Fix Central.

   https://www-945.ibm.com/support/fixcentral/

   ]
2. [

   Run the upgrade tool from the master node in the IBM Spectrum Discover cluster.

   ]

Preparing to run the upgrade tool

Procedure

1. Stop data ingest before you run the upgrade tool.
2. Make sure that you have the most current and up-to-date authentication certificates within IBM Spectrum Discover:

   a) Log in to the IBM Spectrum Discover console with a user ID of *<moadmin>* and a password of *<Passw0rd>*

   b) Run the following command:

```
sudo /etc/cron.hourly/icp_login.sh
```

Running the upgrade tool

Procedure

1. [
   The IBM Spectrum Discover upgrader is a tarball or tarfile. Extract this tarball to the `/opt/ibm/metaocean` directory on the master node of the IBM Spectrum Discover cluster:

   ```
   [moadmin@ metaocean]$ cd /opt/ibm/metaocean
   [moadmin@ metaocean]$ sudo tar xJf sd_upgrader_2012.tar.xz
   ```

   This extracts both new media and new roles for deploying the upgrade.

   ]
2. [
   Use the upgrade command to initiate the upgrade:

   ```
   [moadmin@ metaocean]$ sudo ./upgrade
   ```

   [Running the upgrade tool can take up to several hours, but run time depends on the amount of data that is ingested]

   ]

[Upgrades from releases preceding IBM Spectrum Discover version 2.0.1

About this task

The upgrade process can reboot several times depending on which version of IBM Spectrum Discover is being upgraded:

IBM Spectrum Discover version 2.0.0.2
Upgrades CentOS, ICp, Kafka, Db2wh

IBM Spectrum Discover version 2.0.0.3
Upgrades ICp, Kafka, Db2wh

Remember: Both CentOS and ICp upgrades require reboots.

A reboot disconnect the Secure Shell (SSH) session. However, the upgrader process is displayed immediately upon reconnection after the reboot. The message-of-the-day (MOTD) displays a message indicating that an upgrade is in progress:

```
# BEGIN ANSIBLE MANAGED BLOCK
***********************
* UPGRADE IN PROGRESS *
***********************
# END ANSIBLE MANAGED BLOCK
```

This statement precedes the running upgrader status, whose output comprises the end (or tail) of the upgrader log file data. Use **CTRL + C** to escape without affecting the running upgrader process.

Once the upgrade has completed, the MOTD is replaced with a message that displays the new or upgraded IBM Spectrum Discover version. You do not have to immediately reconnect after a reboot because the upgrader runs automatically as soon as the system returns online. An upgrade can be initiated, and when a new SSH session displays the MOTD without upgrader progress the upgrade has completed.

]

Upgrades from IBM Spectrum Discover version 2.0.1 or later

About this task

Neither CentOS nor ICp is upgraded, so no reboot is required during this upgrade. Upgrade status is displayed in the console where the upgrade command is executed and the upgrade process is stopped if the session is disconnected.

For information on debugging an upgrade, see Debugging a hung upgrade.

Applying the license file

As a prerequisite, you must have the `ibm-spectrum-discover-unrestriced.lic` license file, or an alternative file provided by IBM.

Procedure

1. Log into the machine using SSH.
2. Copy the license file (`ibm-spectrum-discover-unrestriced.lic`) to the IBM Spectrum Discover machine.
3. Get an auth token for the API.

```
TOKEN=$(curl -ks -u sdadmin:<password> https://localhost/auth/v1/token -I | awk '/X-Auth-
Token/ {print $2}')
```

4. Load the license from the license file.

```
LICENSE=$(cat ibm-spectrum-discover-unrestriced.lic)
```

5. Push the license to the server.

```
curl -k -H "Authorization: Bearer ${TOKEN}" -H "Content-Type: application/json"  -X PUT --data
"{\"license\":\"${LICENSE}\"}" https://localhost/api/license/
```

6. To verify the license, go to the administration section of the Web UI, or check the license with the API.

Accessibility features for IBM Spectrum Discover

Accessibility features help users who have a disability, such as restricted mobility or limited vision, to use information technology products successfully.

Accessibility features

The following list includes the major accessibility features in IBM Spectrum Discover:

- Keyboard-only operation
- Interfaces that are commonly used by screen readers
- Keys that are discernible by touch but do not activate just by touching them
- Industry-standard devices for ports and connectors
- The attachment of alternative input and output devices

IBM Knowledge Center, and its related publications, are accessibility-enabled. The accessibility features are described in IBM Knowledge Center (www.ibm.com/support/knowledgecenter).

Keyboard navigation

This product uses standard Microsoft Windows navigation keys.

IBM and accessibility

See the IBM Human Ability and Accessibility Center (www.ibm.com/able) for more information about the commitment that IBM has to accessibility.

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

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

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

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

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

Any references in this information to non-IBM 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.

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

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

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

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

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

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

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

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

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.

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

© (your company name) (year).
Portions of this code are derived from IBM Corp.
Sample Programs. © Copyright IBM Corp. _enter the year or years_.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

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

Intel is a trademark of Intel Corporation or its subsidiaries in the United States and other countries.

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

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

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

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

Terms and conditions for product documentation

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

Applicability

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

Personal use

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

Commercial use

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

Rights

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

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

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

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

IBM Online Privacy Statement

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

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

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

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

Index

Product Number: 5737-I32
5737-SG1

SC27-9601-05