VERITAS

Long-Term Data Retention With Veritas™ NetBackup

Cloud Object Storage

This white paper provides a technical overview of cloud storage as a long-term retention storage solution with Veritas NetBackup™. It highlights the overall solution architecture components, integration flow and best practices

TABLE OF CONTENTS

INTRODUCTION

EXECUTIVE SUMMARY

More companies are venturing to the public cloud as an option to retain data for the long term and/or to safeguard data from on-premises failures, attacks and disasters. Some of their main requirements include cost, security and insight into what data can be sent to the cloud. The Veritas suite of products alleviates some of these issues and concerns. Veritas NetBackup, for instance, includes encryption features for data at rest and in motion for security. It also features storage efficiency technologies such as deduplication and compression to reduce egress and ingress costs to and from the cloud. Veritas Information Studio provides valuable insights to help organizations identify which data should remain on-premises or which could be moved to the cloud.

NetBackup also supports sending and retrieving data to and from tiered layers of object storage with varying costs, availability and performance for long-term retention and preservation of organizations' critical digital assets.

SCOPE

This document is a good primer for customers, partners and Veritas field personnel interested in leveraging cloud object storage for long-term data retention needs using NetBackup. It describes the solution's components, its value and some practical use cases. For installation, configuration and administration of each of the products discussed in this white paper, please refer to the appropriate Veritas product documentation.

SOLUTION VALUE

The Veritas Enterprise Data Services Platform (EDSP) allows you to realize the full potential in your organization's data, ensuring availability, protection and insights across data centers or multiple public clouds, providing a truly infrastructure-agnostic approach. This approach offers the modern enterprise valuable advantages:

- **Deliver insights**—Organizations tend to blindly back up data and not remove stale or orphaned data from their primary storage. Information Studio provides a view into digital assets residing on primary storage and backup storage for help in data placement and throughout the entire data lifecycle. With new regulatory requirements imposed by numerous countries, it's crucial for organizations to understand what data they have as well as its value and liability.
- **Reduce risk**—NetBackup allows organizations to encrypt data at rest and prior to transmitting it to cloud object storage. For data in flight, NetBackup uses the Secure Sockets Layer (SSL) protocol for data transfers between NetBackup and cloud storage for enhanced security.
- **Minimize cost**—With the Veritas Deduplication Engine and compression features, organizations can reduce the amount of data sent or retrieved from cloud storage, minimizing overall costs.

SOLUTION KEY FEATURES

Key features that organizations look for in an off-premises, long-term retention solution include insights, security, storage efficiency and protection options.

INSIGHTS

One of the challenges organizations often face is deciding which data to send to the public cloud. Information Studio has the capability to classify primary sources and/or scan the NetBackup catalog to quickly acquire information about the data on the sources that have been backed up. The ability to identify areas of risk, value and ROT (redundant, obsolete, trivial) improves operational efficiency, reduces storage cost and minimizes risk and liability.

SECURITY

NetBackup has security features that protect all NetBackup components and operations at different security implementation levels such as the data center and the enterprise. Refer to the NetBackup Security and Encryption Guide for further details on NetBackup's security implementations and levels.

For enhanced security, NetBackup also offers encryption of data. Any encryption done by NetBackup is maintained on the cloud storage target. For more information on how NetBackup conducts encryption specifically for cloud storage servers, refer to the NetBackup Cloud Administrators Guide. When using NetBackup deduplication technology, there is encryption for deduplicated data that is separate and different from the NetBackup policy-based encryption. For more information on implementation, refer to the NetBackup Deduplication Guide. Additional security employed for this solution is the requirement to use access keys or identity and access management (IAM) roles when configuring cloud storage with NetBackup. You can also secure data transfers to and from cloud storage providers by enabling the NetBackup SSL feature.

STORAGE EFFICIENCIES

Support for storage efficiency is one of the main factors organizations consider when choosing a long-term retention storage platform solution. The ability to maximize storage space assists in reducing overall cost. Organizations can further space-optimize backup images stored on object storage using the Veritas Deduplication Engine. The Deduplication Engine that powers the media server deduplication pool (MSDP) in NetBackup is an enterprise-class, software-defined storage solution that scales out to any infrastructure while delivering dramatic cost savings with data reduction. Once data is stored in the Deduplication Engine storage format, it becomes highly portable and is optimized for transport to any compatible target on any infrastructure.

NetBackup also supports compression prior to sending data to cloud object storage. Compression improves storage utilization by reducing the number of bits required to represent data. The type of data defines the degree to which a file can be compressed. Data types that compresses well include text files or unstripped binaries. Data that is already compressed and stripped binaries are not good candidates for compression. For detailed information on NetBackup compression attributes, refer to the NetBackup Cloud Administrators Guide and the NetBackup Deduplication Guide.

PROTECTION

In the NetBackup 8.2 release as part of the automated disaster recovery in the cloud, there is a new feature called Image Sharing that extends the capabilities of NetBackup CloudCatalyst. With Image Sharing, NetBackup will not only send deduplicated data to a cloud storage target but also the image metadata. This new feature allows for recovery of images in the cloud when the on-premises NetBackup catalog is not available due to corruption, power outage, network issues etc. Organizations can reconstruct data and metadata in the cloud or a different data center by launching a new CloudCatalyst server and attaching it to existing cloud storage.

SOLUTION ARCHITECTURE OVERVIEW

Figure 1 depicts a high-level overview of the Veritas data protection solution with cloud object storage for long-term retention. An integral part of this solution is the use of Information Studio to provide information on the data residing on-premises to help make decisions on the lifecycle of the data. A report generated by Information Studio assists users in making informed decisions about which data to store on-premises or send to the cloud and/or define the storage lifecycle policies of the data. For instance, the lifecycle of backup data can first reside on-premises on a NetBackup Appliance or BYOS (Build Your Own Server) for the short term, then move to Veritas Access Appliances for mid-term retention and finally to cloud storage for long-term retention and disaster recovery. NetBackup supports different storage classes or access tiers with public cloud storage providers. These classes or tiers of storage differ in terms of cost, usage, restore time, availability and other services.

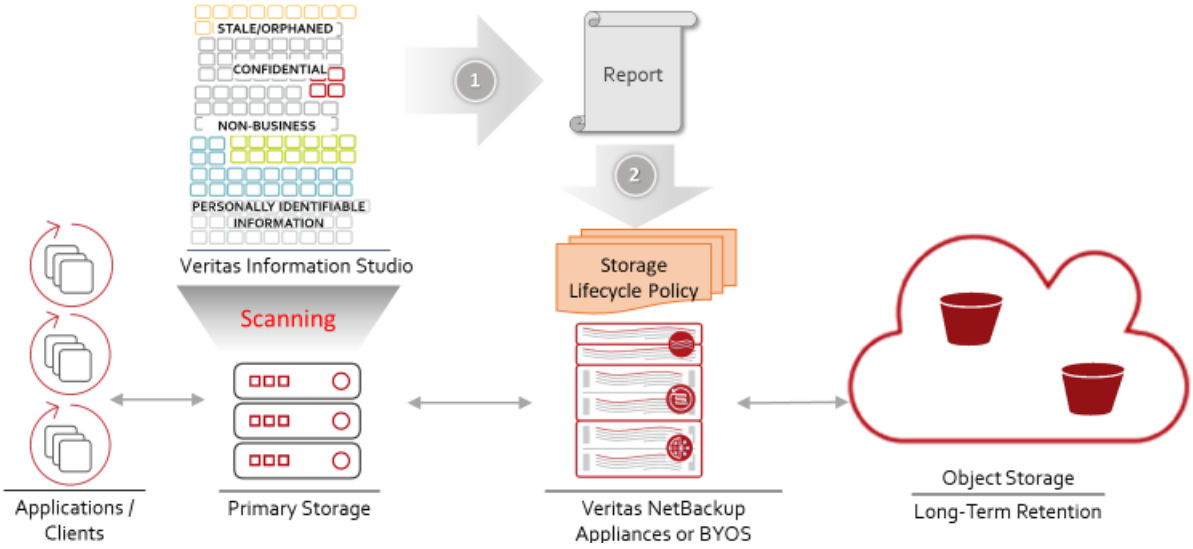


Figure 1. A high-level solution overview of NetBackup with cloud storage.

SOLUTION COMPONENTS

To get a better understanding of how cloud storage services work with NetBackup for long-term retention, we'll go into further detail about all the involved solution components.

INFORMATION STUDIO

Information Studio provides visibility by connecting to varied primary data sources on-premises, including OpenText™, OpenText LiveLink®/Documentum, IBM FileNet, Microsoft® Exchange/SQL Server/SharePoint®, Oracle® database, SMB shares such NetApp, Dell EMC Celerra/VNX/Isilon®, Hitachi and Windows®. It can also scan NetBackup catalogs to harvest information on data that has been backed up. In addition to primary sources, Information Studio can connect to cloud sources such as Microsoft® OneDrive, SharePoint Online, Amazon Web Services (AWS), the Google Cloud platform ad Microsoft Azure. Depending on the data source, it classifies the data into certain categories such as ownership, age, size, activity, stale, non-business, user risk, type and data patterns. This classification enables administrators to identify data that can be archived or tiered to cheaper storage, enforce security and perform information lifecycle management and risk analysis. Architecturally, Information Studio consists of three main components:

- **Hub**—Responsible for the management of jobs, maintenance of configuration information, storage of metadata, logging and generation of reports.

- **Data engine**—Responsible for topology discovery of each data source, scanning and collection of metadata, classification and sending data to the hub.
- **Windows Connector**—Only required when connecting and scanning Windows-based applications and data sources such as Microsoft SQL, Microsoft SharePoint, SMB/CIFS shares and Oracle.

As shown in Figure 2, Information Studio's hub and data engine are containerized services running on top of the Veritas Operating System (VxOS), a customized Linux OS based on Red Hat® Enterprise Linux (RHEL). The containers and VxOS are packaged as an Open Virtual Appliance (OVA) file and deployable on a VMware® hypervisor. The Windows Connector is a separate executable that is run on a Windows host on bare-metal or virtual machines (VMs) and is only required if connecting to a Windows-based data source. Refer to the Veritas Information Studio white paper for more details on the architecture.

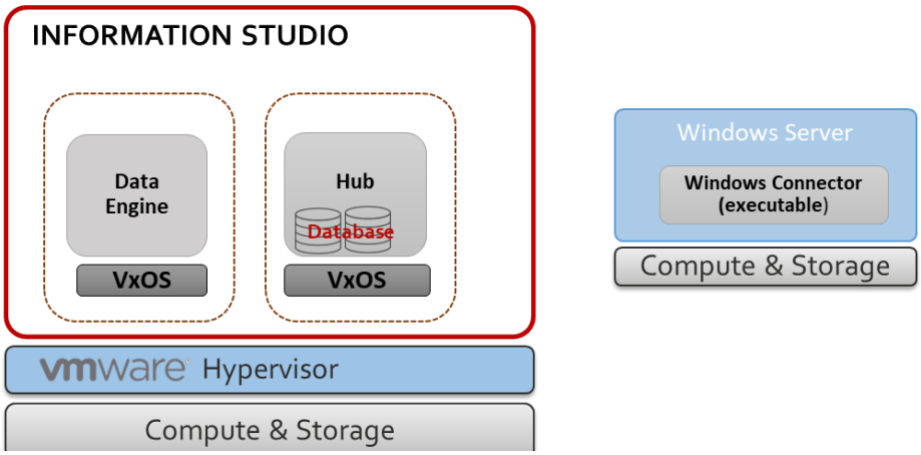


Figure 2. An overview of Information Studio's components.

Depending on the amount of data to be classified, you can deploy a single or multiple instance of the data engine component. For example, if the data source is NetBackup, then only the hub container is required because there is a small-scale data engine running within the hub by default. If the data source is in a remote location or on several SMB shares, you would deploy separate or multiple remote data engine instances. Usually, the data engine is placed in the same geographic region as the data sources being scanned and classified. Minimum VM requirements to run OVA and executables is shown in Table 1.

Table 1. Minimum specifications for Information Studio components.

Veritas Information Studio Components	Minimum Specifications
Hub	16 cores 80 GB RAM Disks: • 500 GB for the Veritas Operating System (VxOS) • 1 TB for data ESX Server version 6 minimum

Remote Data Engine	16 cores 32 GB RAM Disks: • 500 GB for the Veritas Operating System (VxOS) • 500 GB disk for data ESX Server version 6 minimum
Windows Connector	4 cores 8 GB memory Disk: 100 GB free disk space Windows 2012 R2 or Windows 2016 (64-bit versions)

As shown in Figure 3, Information Studio does the following when processing data sources for classification:

1. The data engine or Windows Connector connects to the data sources and conducts a topology discovery. The data engine then pulls the items from the data sources. As previously mentioned, if the data source is Windows-based, then the Windows Connector is part of the communication and data path.
2. The data engine scans to capture the metadata and then classifies the data.
3. The metadata captured during scan and classification is sent to the hub for storing and further querying.
4. Users can query and filter the data using the APIs or web graphical user interface (GUI) based on custom or preconfigured criteria. The results of this filtration are presented within the web GUI and/or a comma-separated variable (CSV) or SQLite file report. The report contains a list of files or items and their associated metadata that meet the user-specifiedfilters.

You can initiate data classification tasks manually or at scheduled intervals (daily, weekly etc.). For details on installation, configuration and deployment, refer to Veritas Information Studio product documentation.

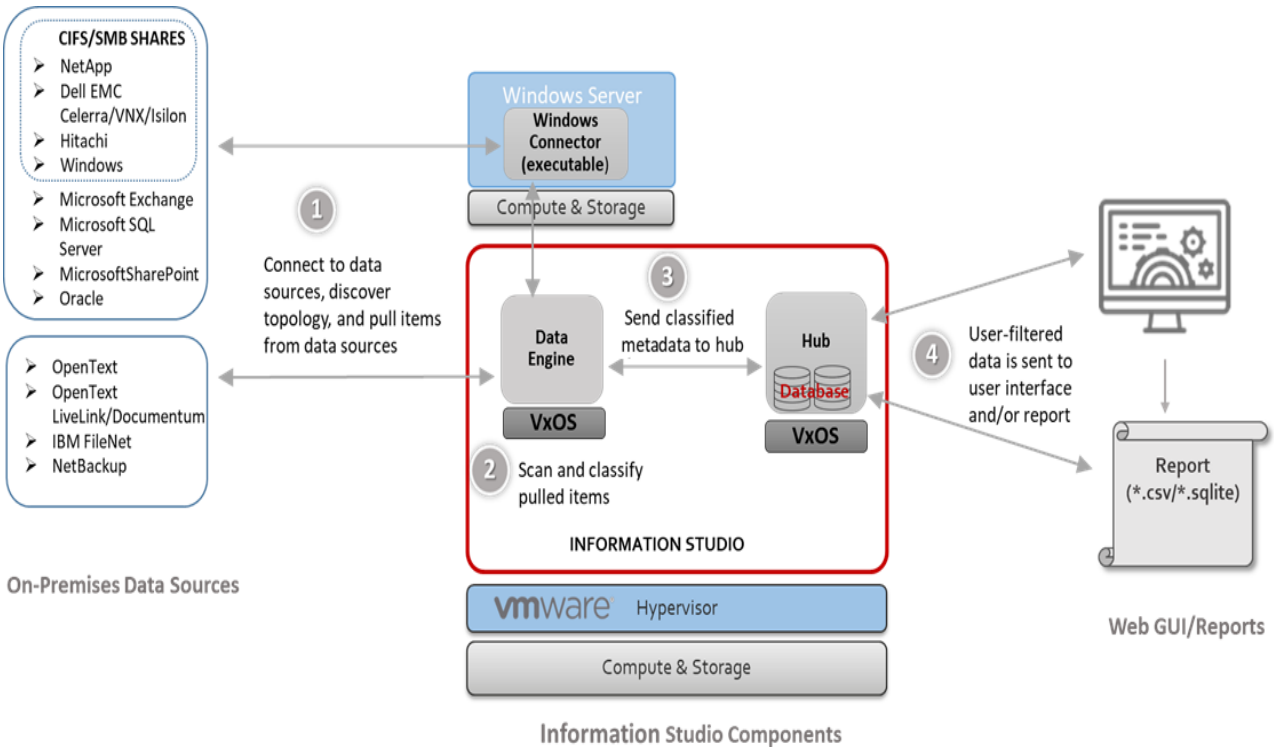


Figure 3. An overview of the Information Studio process flow.

Information Studio is a powerful tool. In addition to custom policy definition where users can find data based on specific pattern or conditions, it has 120+ preconfigured data classification policies and 700 data patterns to identify common data privacy and regulatory compliance principles. Some examples of preconfigured policies include:

- **Corporate compliance**—authentication, "company confidential" and IP, ethics and code of conduct, IP addresses, PCI-DSS and proposals/bids.
- **Financial regulations**—bank account numbers, credit card numbers, GLBA, SOX, SWIFT Codes and U.S. financial forms.
- **Health regulations**—Australia Individual and Canada Healthcare Identifier, IDC 10 CM diagnosis indexes, medical record numbers, U.S. DEA numbers and U.S. HIPAA (Health Insurance Portability Accountability Act of 1996).
- **International regulations**—Australia/Canada/U.K./U.S. driver's license and passport numbers, Australia tax, U.K. Unique Tax Reference (UTR), U.S. Social Security number and Taxpayer ID, Canada SIN, France National ID, Italy Codice Fiscale and Switzerland National ID.
- **Personal identifiable information** (PII) from 35+ countries.
- **Sensitive data policies** from 35+ countries.
- **U.S. state regulations**—criminal history, FCRA, FERPA, FFIECE, FISMA, IRS 1075 or SE.
- **U.S. federal regulations**—California Assembly Bill 1298 (HIPAA), California Financial Information Privacy Act (SB1) and Massachusetts regulation 201 CMR 17.00 (MA 201 CMR 17) and newly evolving policies like the California Consumer Privacy Act (CCPA).

Figure 4 shows an example of Information Studio's web GUI, displaying information such as the amount of stale data, where your data resides, item extensions and age. You can manually inspect the reports generated by Information Studio and use them to define the "Backup Selection List" within NetBackup backup policies. You can use storage lifecycle policies to specify the data that can be kept on-premises and/or duplicated to cloud object storage. Knowing what type of information is stored in data sources

allows organizations to make more informed decisions about what to do with their data for storage optimization, security, compliance, archival and long-term retention.

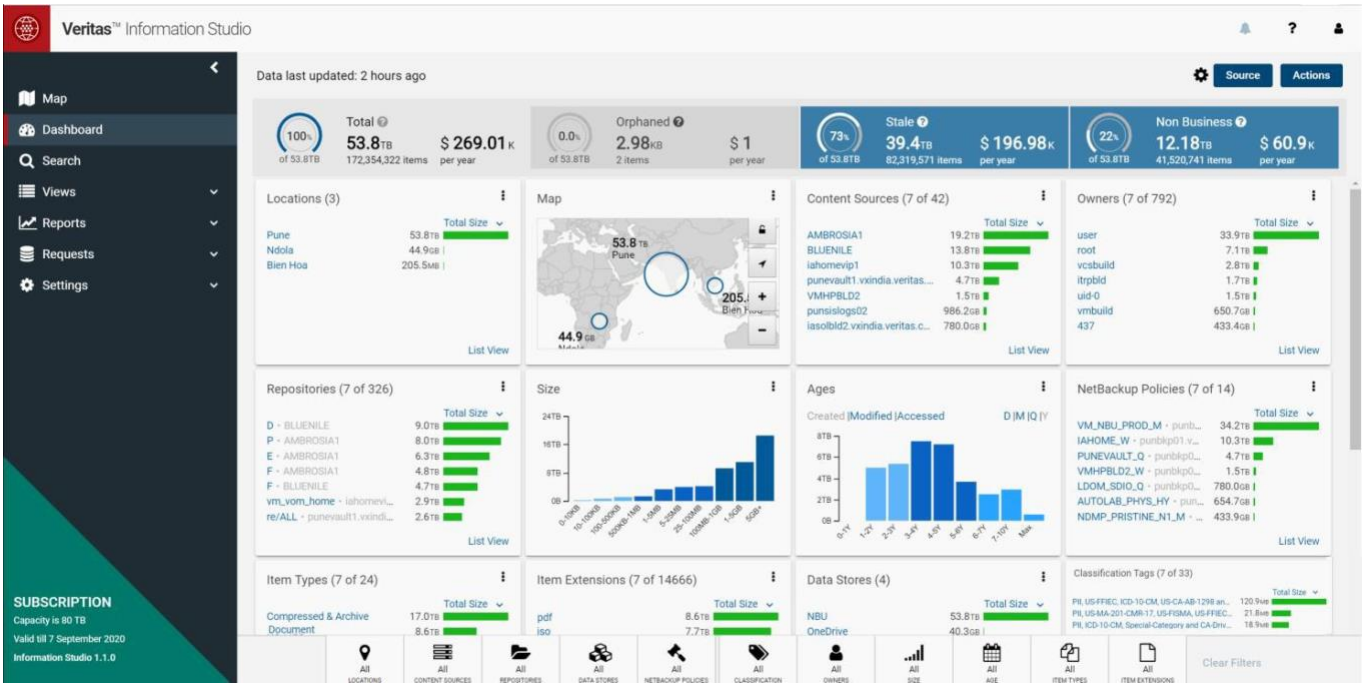


Figure 4. A sample view of Information Studio's web graphical user interface (GUI).

NETBACKUP

NetBackup provides protection for a wide variety of data and platforms such as operating systems, virtual systems, databases and applications, files and a variety of content. It has many features to speed up backups, snapshot management and backup automation and provide insights on where active and inactive backups are located. It has the capability to back up data to tape, storage array network (SAN), network-attached storage (NAS) and public or private clouds. Schedules, retention periods and the ability to tier to different types of storage are defined in policies or storage lifecycle policies (SLP).

A typical NetBackup environment consists of three components:

- **Master server**—Manages and controls backup and recovery activities and hosts the catalog that contains policies and schedules, metadata about backup jobs and media, devices and images.
- **Media server**—Writes client data as backup images to varying types of storage such as local disks, tape, NAS, SAN and cloud and later restores the data to the client as instructed by the master server.
- **Clients**—NetBackup client components are installed on hosts that have the data to be backed up and are responsible for sending and receiving data to and from the media server for backup and recovery.

The master and media server components can be in one system or distributed across multiple servers, depending on the number of clients and backup workload. For a small environment, the master and media server typically can exist in one server, whereas for a large environment, there is usually one dedicated master server and several media servers. NetBackup can also be configured in a multi-domain where there are separate instances of master servers in different locations. In a multi-domain environment, each NetBackup domain is independent but can be centrally managed by NetBackup OpsCenter, a web-based console for managing, monitoring and reporting on NetBackup operations. Figure 5 illustrates a sample configuration of a set of clients with a dedicated master server and one media server.

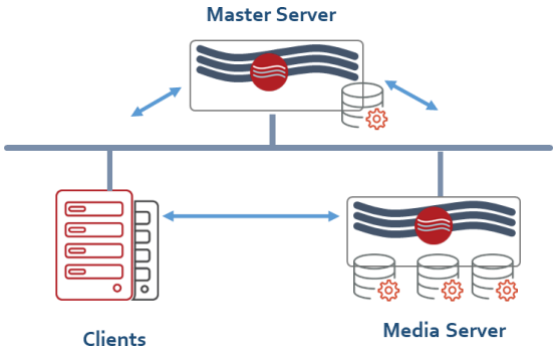


Figure 5. An example of a dedicated NetBackup master server and one media server protecting several clients.

NetBackup is very flexible in its deployment. You can deploy it using a Veritas Appliance solution on commodity servers (also known as Build Your Own Servers or BYOS) or on a mixture of both. You also can run it on bare-metal on-premises, cloud-based VMs or in containers when using the Veritas Flex 5340 Appliance. Following are highlights of each of these deployment options:

- **NetBackup Appliances**—Purpose-built, highly tuned, scalable and resilient integrated appliances for NetBackup components. These appliances address the most demanding backup and recovery requirements of enterprises.
    - NetBackup 5240—Meant for moderate workloads, the 5240 can scale up to 323 TB of storage capacity.
    - NetBackup 5330/NetBackup 5340—The 5330 and 5340 Appliances are for demanding workloads requiring higher usable capacity that can scale up to 1,375 TB and 2,111 TB, respectively. The 5330 and 5340 models have high-availability configurations that include an additional node to continue operations should the active node fail.
    - NetBackup Flex 5340—The Flex 5340 Appliance supports container technology, allowing users to launch multiple containers with different roles of master, media or CloudCatalyst servers in one Appliance. It also supports the creation of multiple domains in one Appliance. The appeal of the Flex Appliance is its multi-tenant capability and the ease of deploying a full NetBackup environment with multiple independent versions of NetBackup quickly.
- **NetBackup Build Your Own Server (BYOS)**—You can deploy NetBackup components on commodity servers that run on a Linux or Windows platform. Refer to the NetBackup Software Compatibility List for a full list of supported platforms. In production, the minimum requirement for a master server is 4 cores and 16 GB of memory. Each media server requires a minimum 4 GB of memory and a minimum of 512 MB for clients. Other than hardware and platform differences, there is also a difference in the maximum MSDP capacity that can be set up on a single server in BYOS. The MSDP capacity for BYOS is limited to 250 TB per server for systems configured with Red Hat Enterprise Linux (RHEL), Windows Server and SUSE Linux and 64 TB for others.
- **NetBackup Virtual Appliances—**You can also run NetBackup on virtual appliances; however, this deployment is mainly appropriate for remote offices. Implementing NetBackup on VMs is a simple deployment and minimizes capital expenditures. Refer to the NetBackup Software Compatibility List for more information on supported hypervisors.
- **NetBackup in the Cloud Marketplace—**NetBackup is available for automated deployment in the Amazon AWS and Microsoft Azure marketplaces. For more information, check the "NetBackup in the Cloud – Deployment Templates" section of the NetBackup Software Compatibility List.

The **Access Appliance** is part of the Veritas appliance portfolio and can serve as an on-premises, mid-term or long-term retention storage target for NetBackup. For organizations seeking to extend their on-premises disk-based storage platform for faster recovery times, control and/or simplicity, the Access Appliance is a turnkey storage solution designed for high capacity and cost optimization. The Access 3340 Appliance is composed of two clustered compute nodes and one primary storage shelf and up to three additional expansion storage shelves. It can scale up to 2,800 TB of usable space.

There are two ways to store backup images in a public cloud from NetBackup:

1. **Deduplication** (optimized duplication)—Deduplication using NetBackup MSDP deduplication technology.
2. **Without deduplication** (traditional duplication)—Backup images are duplicated to the public cloud from NetBackup.

Deduplication

Backup images are generally ideal for deduplication because the probability of encountering duplicated blocks of data is higher when compared to other data types such as encrypted data. The deduplication ratio defines how well data can be deduplicated. The higher the ratio, the more space deduplication saves. Deciding on whether to deduplicate your data or not depends on several factors: data type, data change rate, retention period and backup policy. For instance, encrypted data is inherently unique and may not benefit from deduplication savings. Likewise, data with a high change rate will not take advantage of the savings long enough to justify the overhead imposed by deduplication. In the context of backup images, daily full backups will have higher deduplication ratios when compared with incremental or differential backups.

NetBackup MSDP is a proprietary Veritas deduplication technology. If the data is first placed in an MSDP and then duplicated to another storage platform that does not support MSDP technology, NetBackup would rehydrate deduplicated data prior to sending data to the storage platform. Rehydration involves putting the backup image back to a non-deduplicated form. NetBackup allows for inline deduplication of backup images on either the client or media server. The difference between the client-side or media server deduplication is where the deduplication occurs. For a client-side deduplication target, the backup data is first deduplicated on the client before being sent to the target storage. Client-side deduplication uses available resources on clients and reduces network traffic because deduplicated data is sent over the network. In either scenario, the backup images are placed in an MSDP.

Architecturally, NetBackup MSDP deduplication is composed of the following main components:

- **Deduplication Plug-in**—Separates the data into segments or chunks. The plug-in uses a hash algorithm to calculate fingerprints to identify each unique segment and compares incoming data fingerprints with the fingerprints of existing data.
- **Deduplication Engine** (spoold)—Manages and stores the fingerprint database and metadata, stores unique segments or uses a reference or pointer to the data already stored and conducts integrity checks.
- **Deduplication Manager** (spad)—Maintains the configuration, controls and dispatches the internal processes, security and events handling.

NetBackup MSDP uses SHA-2 (SHA256) for the hash algorithm. The chunk segment size unit used to compute fingerprints is by default a fixed length of 128 KB or configurable to a variable-length size based on the chunk boundary. NetBackup MSDP also compresses deduplicated data for further storage efficiency. Furthermore, there is an option to encrypt deduplicated data. Both compression and encryption (if enabled) are performed after the fingerprint is calculated and prior to sending data to the target storage. For more information on the architecture of NetBackup MSDP deduplication technology, refer to the NetBackup Deduplication Guide.

CloudCatalyst is a crucial component in sending deduplicated data to cloud storage without rehydration. CloudCatalyst can be deployed on a dedicated bare-metal host, provisioned as a NetBackup 5240 Model G appliance or as a container in a Flex Appliance. You can also deploy it as a VM running on-premises or in the cloud. The CloudCatalyst Appliance is essentially the NetBackup 5240 Appliance with the G option configured to have more memory (192 GB). Note that there is a separate SKU for the CloudCatalyst Appliance. You can implement multiple CloudCatalyst instances; however, each CloudCatalyst instance can only write to one on-premises or public cloud storage platform and only to one bucket (1 PB maximum qualified). A BYOS deployment of CloudCatalyst requires RHEL 7.3 or higher, and the minimum amount of memory and disk cache is configurable and dependent on amount of data

being backed up.

Traditional Duplication (without deduplication)

In some cases, deduplication of backup images is not ideal. Backups that have a strict time limit for restores, have a high rate of change or are encrypted are not good candidates for deduplication. For these types of data or backups, it's best to send images to the public cloud without deduplication. Data sent to cloud storage from NetBackup can be tagged to use a storage class or access tier based on the functionality available from the cloud provider.

SOLUTION INTEGRATION FLOW

All data goes through a lifecycle from being created, read, modified, moved to other tiers of storage and eventually expired or deleted once it's no longer of use. When data is actively used, it resides in primary storage and is backed up to secondary storage for protection. When data or backup data is infrequently accessed, organizations move it to cheaper storage on-premises and/or off-premises. Understanding data in terms of usage, age, type and whether it contains PII, is non-business or is subject to regulatory compliance is crucial in determining where to move that data across storage platforms or different tiers of storage on-premises or off-premises.

NetBackup manages the lifecycle of data using storage lifecycle policies (SLPs). NetBackup backup policies and/or SLPs define the path or flow of data. You can define backup policies to send data either to a single target or to an SLP. An SLP defines the lifecycle objectives of the data from backup to duplication to varying storage types and/or replication to different domains. Data is sent to a cloud storage using one of the built-in cloud connectors deployed on NetBackup media servers. The solution integration flows described in this section includes:

- Identification of data to send to the public cloud using Information Studio.
- Optimized duplication (deduplication) to the cloud using CloudCatalyst.
- Optimized duplication (deduplication) to the cloud from an Access Appliance.
- Traditional duplication to the cloud.

IDENTIFICATION OF DATA TO SEND TO THE PUBLIC CLOUD

You can identify what data to send to the public cloud via visual inspection or use Information Studio, if applicable, to do a more granular and user-defined filtered search to identify appropriate data. As shown in Figure 6, Information Studio scans items within data sources such as SMB shares or Microsoft SharePoint and classifies them based on certain preconfigured filters or user-defined criteria such as PII, non-business and activity. Information Studio then generates a report either via the web GUI or through APIs.
NOTE: *If the data source is the NetBackup catalog, Information Studio performs no classification; however, it harvests certain attributes from the catalog such as age and file type.*

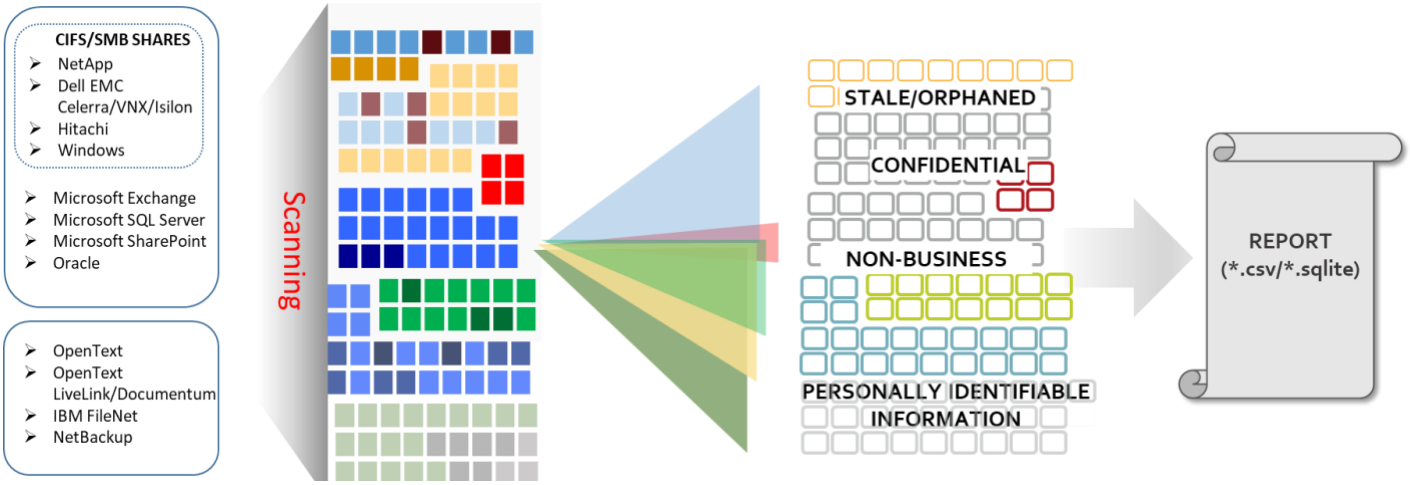


Figure 6. Information Studio's report generation flow.

If you generated the report using the GUI, Information Studio creates a comma-separated variable (CSV) file, and if you used an API to generate the report, it generates a file in either CSV or SQLite format. Some organizations may opt to generate reports in SQLite format to conduct more advanced SQL queries and generate a list of items to send to the cloud based on these queries. Figure 7 shows a sample of the output in CSV format. The contents of this report include a header that describes the type of data in the report such as name, owner, extension, size, count and classification tags. The extracted data is in the subsequent lines.

*Figure 7. A sample snippet of an Information Studio *.csv report.*

```
$ cat nbu.csv | more

Name,Owner,Extension,Size,Count,NetBackupPolicies,MasterServer,CreatedTime,ModifiedTime,AccessedTime,Repository,Co
ntentSource,DataStore,Path,Location,ClassificationTags,ClassifiedTime,PeopleTags,PlacesTags,OrganizationTags,LastNerSca
nTime,LastContentMatchedTime

ArchiveMigration.pdb,user,pdb,65866752,1,VM_NBU_PROD_M,"xxxxx.xxxxx.veritas.com",2013-09-16T07:42:24Z,2013-09-
16T07:40:09Z,2013-09-16T07:40:08Z,F,VM_NBU_PROD_M,NBU,"/F/DFS1/From
HROUS/Products/BE/COLUMBUS/1375D/exe.o/i386/ArchiveMigration.pdb",Pune,,,,,,,

"int8-exp-three-digits.out",uid-26,out,28578,1,VM_SDIO_CICD_Q,"xxxxx.xxxxx.veritas.com",2016-12-16T07:16:59Z,2012-02-
23T22:59:21Z,2012-02-23T22:59:21Z,usr,VM_SDIO_CICD_Q,NBU,"/usr/lib64/pgsql/test/regress/expected/int8-exp-three-
digits.out",Pune,,,,,,,
```

As shown in Figure 8, you would need to further filter this report to pull the file paths and then manually enter them into the NetBackup backup policy "Backup Selection List" attribute using the NetBackup administration console. Alternatively, you could also create an "inclusion" script or "exclusion" list file and feed it into the NetBackup commands using bpplinclude and bpsetconfig, respectively. (We'll look at example scripts that incorporate these commands to populate the "Backup Selection List" in the backup policy later in this white paper.) Once you've defined the backup policy, it can be targeted to an SLP to first do a backup to a local storage target followed by a duplication to a cloud target. Parameters in the SLP determine at what point the second copy is made as well as the retention policies for all the copies in the SLP. The backup policy attribute "Policy Storage" is modified to use the desired storage lifecycle policy.

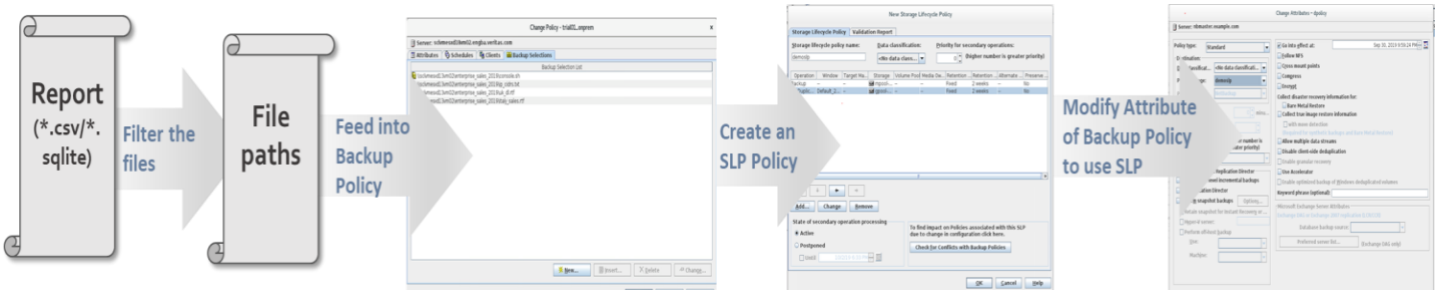


Figure 8. Information Studio's integration flow with NetBackup.

OPTIMIZED DUPLICATION (DEDUPLICATION) DATA FLOW TO CLOUD OBJECT STORAGE

NetBackup CloudCatalyst is required when sending deduplicated data to any of the cloud object storage providers Veritas supports. The most common path to cloud storage is when data from clients is initially deduplicated, written to MSDP storage for short-term retention and subsequently duplicated to cloud storage for long-term retention. You can restore the data from any of the copies that reside on-premises or are retrieved from the public cloud.

Figure 9 illustrates an example where client data is written to an MSDP for short- or mid-term retention (copy 1). This deduplicated data is sent to CloudCatalyst, which caches the data for performance and then uploads the unique data to cloud storage. It's best to do the deduplication on a media server instead of the CloudCatalyst server. Having a separate media server do the backup and deduplication allows CloudCatalyst to use all its resources mainly for caching and transferring the data to cloud storage.

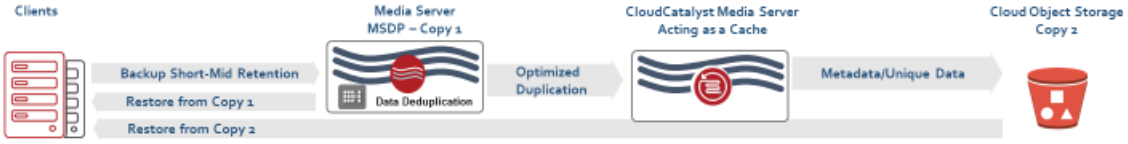


Figure 9. NetBackup data flow from client to media server to object storage using the CloudCatalyst media server.

For restores, data goes through a similar path, but in the reverse direction. By default, if the data is still in the media server, then it's restored from the media server. However, if you're restoring data from cloud storage, the CloudCatalyst cache is first checked for the needed data. If it exists in the cache, data is restored from the cache copy. If data is not present in the CloudCatalyst cache, the data is retrieved from cloud storage, passed to CloudCatalyst and then onto the client for restore purposes.

ACCESS APPLIANCE OPTIMIZED DUPLICATION (DEDUPLICATION) DATA FLOW TO CLOUD OBJECT STORAGE

In some scenarios, the Access Appliance is in the data path for on-premises mid-term or long-term retention prior to sending off-premises. When using Access data deduplication with NetBackup, you can duplicate data to the cloud using NetBackup SLP policies and CloudCatalyst. The SLP would specify to duplicate the data from the Access Appliance to CloudCatalyst to send the data to cloud storage. Figure 10 provides a view of this approach, sending data to the cloud from the Access host running the Deduplication Engine. In this example, deduplicated data is sent to the Access Appliance from the media server, which then does an optimized duplication to cloud storage via CloudCatalyst. The role of the media server during the optimized duplication to the cloud is to control and orchestrate the transfer between the Access Appliance and CloudCatalyst. The actual I/O is between the Access Appliance and CloudCatalyst. You can do a restore either from the Access Appliance (copy 2) or from cloud storage (copy 3). By default, copy 1 is used for restores unless you specify restoring from different copies.

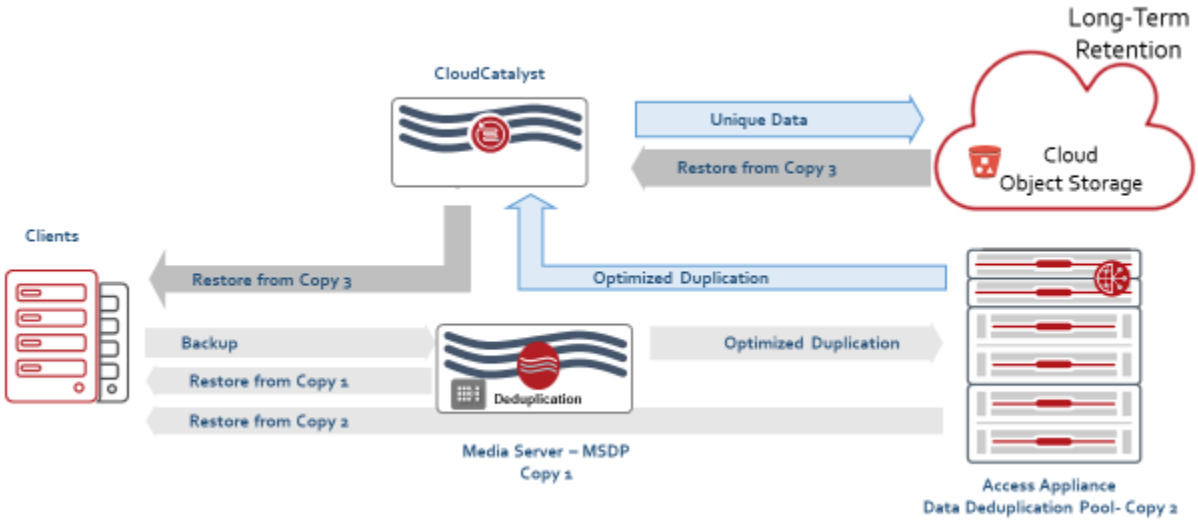


Figure 10. The Access Appliance data deduplication path to cloud storage via CloudCatalyst.

TRADITIONAL DUPLICATION DATA FLOW

For data that doesn't benefit from deduplication, backup data is duplicated from the media server to cloud object storage using the NetBackup OST cloud plug-in installed by default on media servers. The traditional flow for sending data to cloud object storage is for the data to be initially stored on NetBackup AdvancedDisk pool (copy 1) on a media server for short- to mid-term retention and

then copied to cloud storage for long-term retention. You can do restores from either the advanced disk (copy 1, the default) or from cloud storage (copy 2), as shown in Figure 11.

Figure 11. Traditional duplication from NetBackup AdvancedDisk to cloud object storage without deduplication.

BEST PRACTICES AND RECOMMENDATIONS

This section highlights some best practices when using cloud storage as a long-term retention solution for NetBackup to create an optimum deployment.

PRIVACY LAWS

Certain countries have employed the EU's General Data Protection Regulation (GDPR) and privacy laws specifically related to data stored in the public cloud. Although NetBackup stores data in a backup image format, you can still peruse or scan the data for PII, confidential and business-critical information. Use tools such as Information Studio to decide what information is best to store on-premises versus in the public cloud to reduce liability and potential violation of these laws. Also be aware of restoring data in the cloud when using NetBackup automated disaster recovery, "self-describing images." For example, if data is backed up in a Europe region, the NetBackup cloud instances and data restored in the cloud should be in the same region or adhere to your company's GDPR policies. Use of NetBackup encryption would also assist in addressing some of the concerns relating to privacy.

COMPRESSION

For better storage utilization, using NetBackup compression might be an option when deduplication is not ideal and the data type being backed up is compressible. Although compression can reduce the size of a backup, it can also consume server resources. As a best practice, size the media server appropriately for compression. For detailed information on NetBackup compression attributes and considerations, refer to the NetBackup Administration Guide, Volume I. For information on compression for cloud storage targets and deduplication, refer to the NetBackup Cloud Administrators Guide and the NetBackup Deduplication Guide, respectively.

DEDUPLICATION

We recommend you use CloudCatalyst solely for caching and transferring data to cloud object storage and use a separate media server to perform deduplication at the initial backup, especially when handling large amounts of data. If more than 250 TB of MSDP is required, use the NetBackup and CloudCatalyst Appliances as opposed to the BYOS version. NetBackup BYOS has a limitation of 250 TB for the size of MSDP, whereas the size of MSDP on an appliance can be up to 323 TB or 2,111 TB, depending on the model. The maximum MSDP capacity depends on the NetBackup Appliance used as a media server.

When the first copy resides within an MSDP and is then duplicated to the cloud without CloudCatalyst, the data is rehydrated prior to being sent to cloud object storage. Rehydration will increase the time and network resources needed to send data to the cloud. As a best practice, we recommend using CloudCatalyst to send deduplicated data to the cloud; if deduplicated data is not required,

do not place data within an MSDP but instead in other media types such as AdvancedDisk.

NETBACKUP ATTRIBUTES

There are several NetBackup attributes that are best to use when sending or restoring data from cloud storage that supports access tiers or where stored objects can be classified based on retrieval times. For instance, it's best to use the option True Image Recovery whenever possible. Enabling this option significantly reduces the retrieval or restore of data from hours to minutes in some cases. True Image Recovery keeps track of all file system data at the time of backup, so it only restores files that were present at the time of the last backup versus restoring all files.

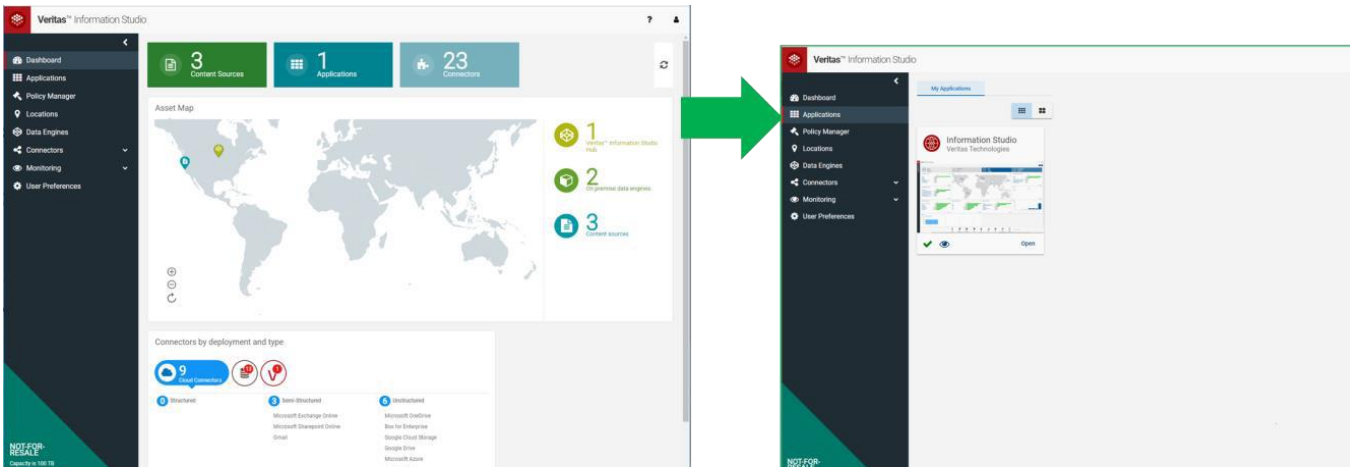
NetBackup also employs an attribute called RETRIEVAL RETENTION PERIOD when used in conjunction with tiered object storage. Set this value within the NetBackup storage server to a minimum of 3 days. Doing so allows a copy of the data to be temporarily reclassified within a holding area for scenarios where it might take several retries to pull back all the necessary data.

INFORMATION STUDIO

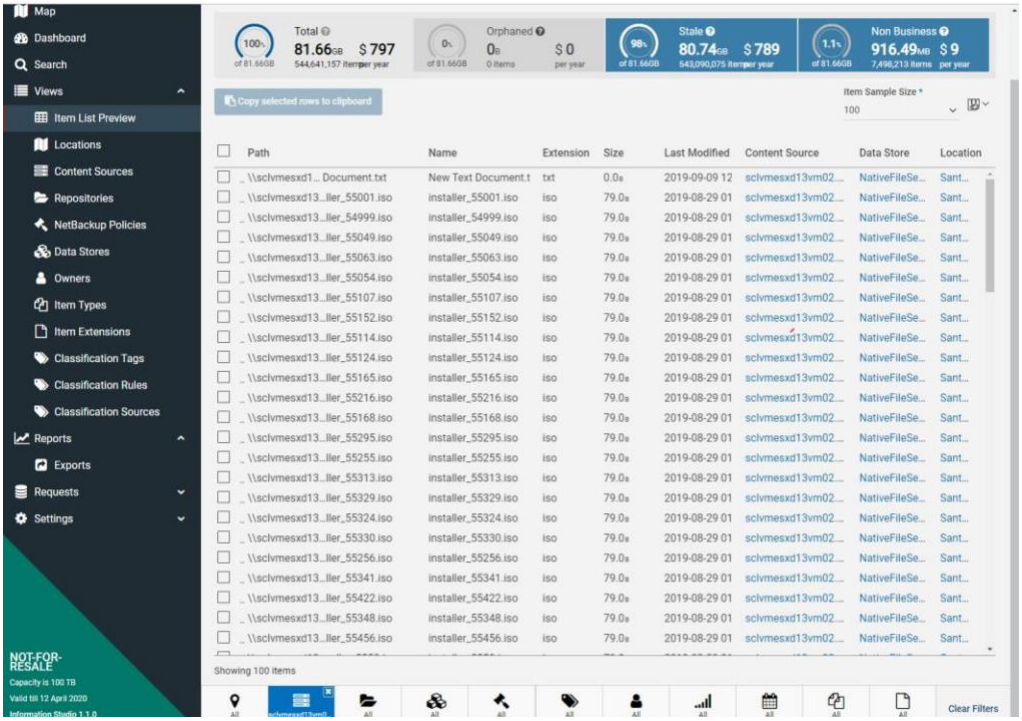
As mentioned previously, you can generate reports using Information Studio's web GUI or via APIs. Below are the steps for generating a report using the GUI.

GENERATE CSV REPORT

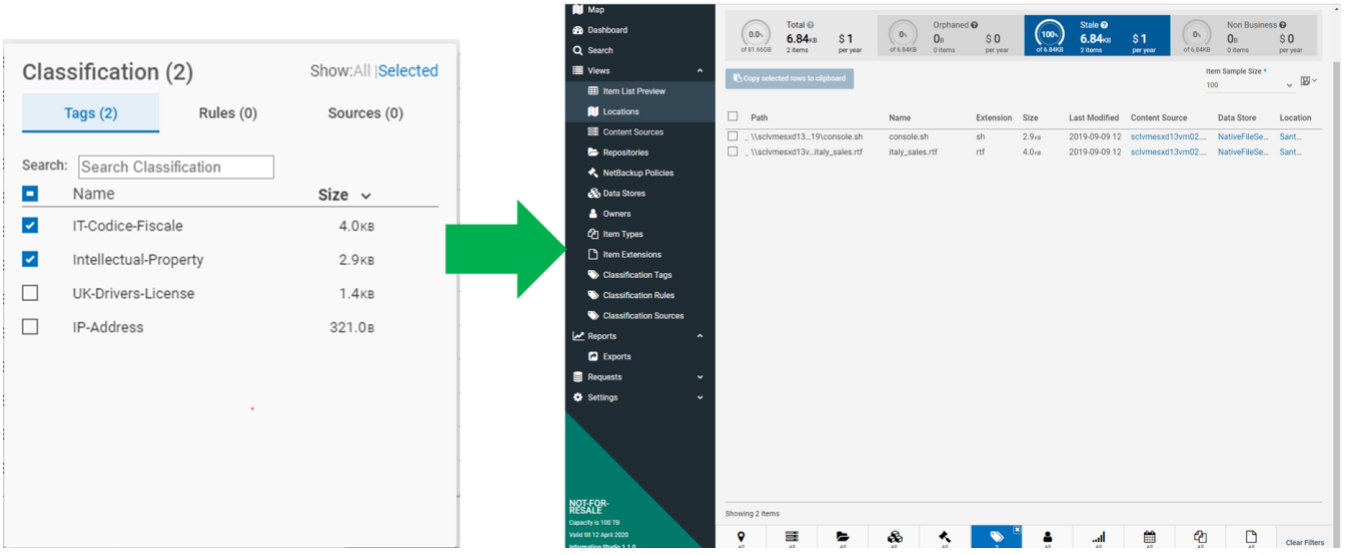
1. Log onto the Information Studio GUI. Click on **Applications** and then click on **Open**.



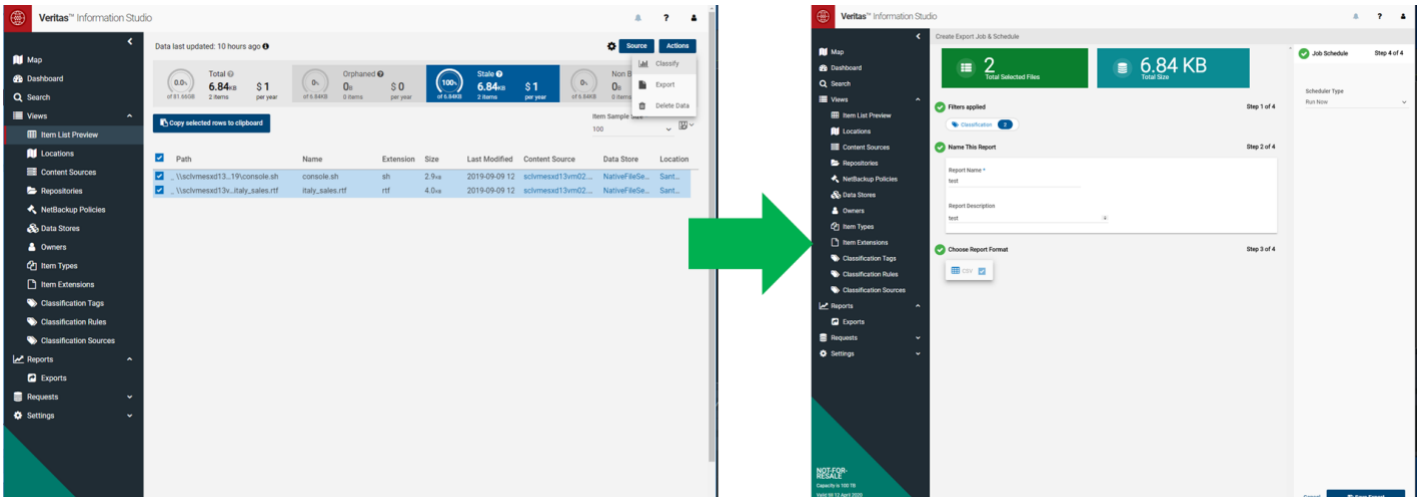
2. Expand **Views** and click on **Item List Preview**. At the bottom are options to select the repositories, data stores, classification, owners, size etc. In this example, from all the data sources, filter based on the predefined classification.



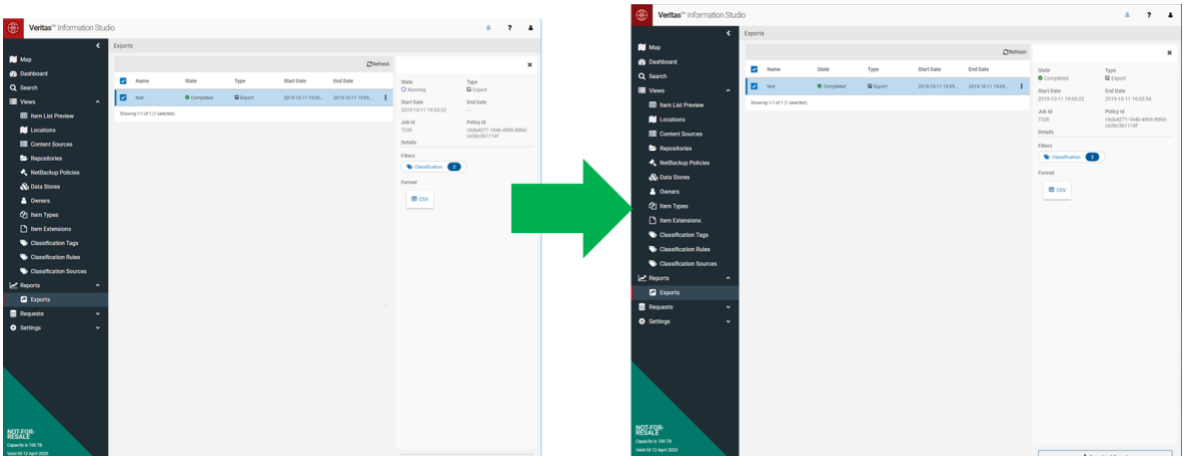
3. Click on **Classifications** and select the predefined classification filters **IT-Codice-Fiscale** and **Intellectual Property**. A set of results will appear on the right pane as shown below. **NOTE**: *The preview of items in the GUI is limited to 1,000 lines; however, when the report is generated, it will contain all filtered data.*



4. Generate the report (*.csv) by clicking on **Actions** at the top and selecting **Export**. Then enter the **name** of the report and a description. Select the scheduler type to be **"Run Now."**



5. Click on **Reports** on the left-side pane to view the progress. Once done, click on **Download Report** to download a *.zip file.



6. Extract the file from the zip file. A sample view of report is shown below. The file paths are highlighted.

```
[root@nbumaster]# cat testout_0.csv
Name,Owner,Extension,Size,Count,NetBackupPolicies,MasterServer,CreatedTime,ModifiedTime,AccessedTime,Repository,
ContentSource,DataStore,Path,Location,ClassificationTags,ClassifiedTime,PeopleTags,PlacesTags,OrganizationTags,LastNe
rScanTime,LastContentMatchedTime
italy_sales.rtf,Administrators,rtf,4046,1,,,2019-09-09T19:13:23Z,2019-09-09T19:13:39Z,2019-09-
09T19:13:23Z,enterprise_sales_2019,"nbuclient.com",NativeFileServer,"\\nbuclient.com\enterprise_sales_2019\italy_sales.rtf"
,"Santa Clara CA","IT-Codice-Fiscale:'Date of Birth','Italy Codice Fiscale'
",2019-09-11T21:53:23Z,,,,,
ip_cidrs.txt,Administrators,txt,321,1,,,2019-09-09T19:14:06Z,2019-09-09T19:14:25Z,2019-09-
09T19:14:06Z,enterprise_sales_2019,"nbuclient.com",NativeFileServer,"\\nbuclient.com\enterprise_sales_2019\ip_cidrs.txt","
Santa Clara CA","IP-Address:'IPv6 Addresses'
",2019-09-11T21:53:23Z,,,,,
console.sh,Administrators,sh,2962,1,,,2019-09-09T19:16:39Z,2019-09-09T19:16:39Z,2019-09-
09T19:16:39Z,enterprise_sales_2019,"nbuclient.com",NativeFileServer,"\\nbuclient.com\enterprise_sales_2019\console.sh","S
anta Clara CA","Intellectual-Property:'Confidential - Generic','Programming Language Source Code'
",2019-09-11T21:53:23Z,,,,,
uk_dl.rtf,Administrators,rtf,1406,1,,,2019-09-09T19:12:38Z,2019-09-09T19:12:52Z,2019-09-
09T19:12:38Z,enterprise_sales_2019,"nbuclient.com",NativeFileServer,"\\nbuclient.com\enterprise_sales_2019\uk_dl.rtf","Sa
nta Clara CA","UK-Drivers-License:'U.K. Drivers License Number'
",2019-09-11T21:53:23Z,,,,,
_sales.rtf","","0","nbuclient.com","4046","italy_sales.rtf","20190909T191323+0000","b3e73199-d348-11e9-b0ab-
215a5907467d","20190909T191339+0000","fbe1cc3e-0db7-4ecc-acbf-1676dbcf089c","false"
```

EXTRACT THE FILE PATHS FROM THE REPORT

You can develop scripts, awk/sed or other tools to extract the file paths from the report generated in the previous section. Below is a sample script that would extract paths and create file paths NetBackup command tools can use as inputs.

1. Extract the file paths from the *.csv report you downloaded in the previous section. The sample script below will extract the files from this output and put it in a form you can enter manually into the "Backup Selection List" of the policy or feed into a script that would create or modify an existing script. ***Usage****: ./getfiles.py [csv_file]*

```
[root@nbumaster]# ./getfile2.py testout_0.csv
Exclude = \\nbuclient.com\enterprise_sales_2019\console.sh
Exclude = \\nbuclient.com\enterprise_sales_2019\ip_cidrs.txt
Exclude = \\nbuclient.com\enterprise_sales_2019\uk_dl.rtf
Exclude = \\nbuclient.com \enterprise_sales_2019\italy_sales.rtf
```

2. When the script is run, it will extract the file paths as shown below.

```
[root@nbumaster]# cat getfiles.py

#!/usr/bin/env python
from csv import DictReader
import sys
def read_command_line_arguments():
if len(sys.argv) != 2:
print_usage() exit()
def print_usage():
print ("Usage: %s" % (sys.argv[0]) + " [csv_file]" )
sys.exit()
read_command_line_argument
s() csv_inputfile = (sys.argv[1])

#Extract filenames from Information Studio Report
with open(csv_inputfile) as file:
csv_reader = DictReader(file)
for row in csv_reader:

##Format the list as per Netbackup exclusion list syntax:

print("Exclude = " + row['Path'])
# Use the below line if do not want to add "Exclude" in output
# print(row['Path'])
```

ENTER THE CLIENT INFORMATION AND FILE PATHS INTO NETBACKUP

You can manually enter the extracted client information and file paths into the backup policy. In this example, the client is a Windows box and the file paths are fed to an exclude list so these files are backed up to the cloud. We've also provided a sample script to modify the exclude list on a client.

1. Redirect the output of the shell script to a *.txt file to feed into the command "bpsetconfig" to exclude it.

```
[root@nbumaster]# cat win_exclude.txt
Exclude = \\nbuclient.com\enterprise_sales_2019\console.sh
Exclude = \\nbuclient.com\enterprise_sales_2019\ip_cidrs.txt
Exclude = \\nbuclient.com\enterprise_sales_2019\uk_dl.rtf
Exclude = \\nbuclient.com\enterprise_sales_2019\italy_sales.rtf
```

2. Run bpsetconfig on the master to set the exclusion on a client machine nbuclient.com, and then bpgetconfig to confirm the exclusion has been configured.

```
root@nbumaster]#
/usr/openv/netbackup/bin/admincmd/bpsetconfig -h nbuclient.com win_exclude.txt

root@nbumaster]# /usr/openv/netbackup/bin/admincmd/bpgetconfig -M nbuclient.com | grep '^Exclude'
Exclude = \\nbuclient.com\enterprise_sales_2019\console.sh
Exclude = \\nbuclient.com\enterprise_sales_2019\ip_cidrs.txt
Exclude = \\nbuclient.com\enterprise_sales_2019\uk_dl.rtf
Exclude = \\nbuclient.com\enterprise_sales_2019\italy_sales.rtf
```

3. The example script shows how you can create a NetBackup policy to use the report generated from Information Studio and conduct a backup. The getfiles.py script used in the previous step was modified to NOT add the word "Exclude" and was used to generate an output TEMPFILE (for example, /root/demo/tmpfile.out) of files to be added to "Backup Selection List" of policy to conduct a backup. The policy created is for one client and a single SMB share and uses the NetBackup commands such as bppolicynew, bpplist, bpplsched, bpplinclude and bpplclients.

```
[root@nbumaster]# cat create_nbu_policy.sh
echo -e "Creating NBU Policy for classified Files using on-prem storage unit(tier)\n"
CIFS_SHARE=enterprise_sales_2019
NEW_POLICY_NAME=demotest1
SCHEDULE_NAME=mysched01a
NBU_MASTER=nbumaster.com
NBU_CLIENT_HOST=nbuclient
ON_PREM_STU=stu001
CSV_REPORT=/root/demo/testout_0.csv
TEMPFILE=/root/demo/tmpfile.out
export PATH=$PATH:/usr/openv/netbackup/bin/admincmd

bppolicynew $NEW_POLICY_NAME -M $NBU_MASTER
bpplist $NEW_POLICY_NAME -U
bpplsched $NEW_POLICY_NAME -add $SCHEDULE_NAME -st FULL -residence $ON_PREM_STU -window 82800 3600
bpplsched $NEW_POLICY_NAME -U
bpplinclude $NEW_POLICY_NAME -delete -M $NBU_MASTER -f "C:\enterprise_sales_2019"

/root/demo/getfiles.py $CSV_REPORT > $TEMPFILE

for file in `cat $TEMPFILE`
  do
    bpplinclude $NEW_POLICY_NAME -add $file
  done

bpplist $NEW_POLICY_NAME -U
bpplclients
bpplclients $NEW_POLICY_NAME -M $NBU_MASTER -add $NBU_CLIENT_HOST Windows-x64 MS-Windows
bpplist $NEW_POLICY_NAME -U |grep HW

echo "Done!!"
```

4. Below is an output of running this script.

```
[root@nbumaster]# sh -x create_nbu_policy.sh
+ echo -e 'Creating NBU Policy for classified Files using on-prem storage unit(tier)\n'
Creating NBU Policy for classified Files using on-prem storage unit(tier)

+ CIFS_SHARE=enterprise_sales_2019
+ NEW_POLICY_NAME=demotest1
+ SCHEDULE_NAME=myschedo1a
+ NBU_MASTER=nbumaster.com
+ NBU_CLIENT_HOST=nbuclient.com
+ ON_PREM_STU=stu001
+ CSV_REPORT=/root/demo/testout_0.csv
+ TEMPFILE=/root/demo/tmpfile.out
+ export PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/root/bin:/usr/openv/netbackup/bin/admincmd
+ PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/root/bin:/usr/openv/netbackup/bin/admincmd
+ bppolicynew demotest1 -M nbumaster.com
+ bppllist demotest1 -U
------------------------------------------------------------------------

Policy Name:      demotest1

  Policy Type:        Standard
  Active:             yes
  Effective date:     10/15/2019 16:33:33
  Client Compress:    no
  Follow NFS Mounts: no
  Cross Mount Points: no
  Collect TIR info: no
  Block Incremental: no
  Mult. Data Streams: no
  Client Encrypt:     no
  Checkpoint:         no
  Policy Priority:    0
  Max Jobs/Policy:    Unlimited
  Disaster Recovery:  0
  Collect BMR info: no
  Residence:          (specific storage unit not required)
  Volume Pool:        NetBackup
  Server Group:       *ANY*
```

```
  Keyword:          (none specified)
  Data Classification:      -
  Residence is Storage Lifecycle Policy: no
  Application Discovery:     no
  Discovery Lifetime:     0 seconds
  ASC Application and attributes: (none defined)

  Granular Restore Info: no
  Ignore Client Direct: no
  Use Accelerator: no

  Clients:     (none defined)

  Include:     (none defined)

  Schedule:    (none defined)
+ bpplsched demotest1 -add mysched01a -st FULL -residence stu001 -window 82800 3600
+ bpplsched demotest1 -U

Schedule:          mysched01a
  Type:            Full Backup
  Frequency:         every 7 days
  Excluded Dates----------
     No specific exclude dates entered
     No exclude days of week entered
  Synthetic:         0
  Checksum Change Detection: 0
  PFI Recovery:      0
  Maximum MPX:       1
  Retention Level:    1 (2 weeks)
  Number Copies:      1
  Fail on Error:      0
  Residence:         stu001
  Volume Pool: (same as policy volume pool)
  Server Group: (same as specified for policy)
  Residence is Storage Lifecycle Policy: 0
  Daily Windows:
```

```
      Sunday    23:00:00 --> Sunday    24:00:00
      Monday    23:00:00 --> Monday    24:00:00
      Tuesday 23:00:00 --> Tuesday 24:00:00
      Wednesday 23:00:00 --> Wednesday 24:00:00
      Thursday 23:00:00 --> Thursday 24:00:00
      Friday    23:00:00 --> Friday    24:00:00
      Saturday 23:00:00 --> Saturday 24:00:00
+ bpplinclude demotest1 -delete -M nbumaster.com -f 'C:\enterprise_sales_2019'
+ /root/demo/getfiles.py /root/demo/testout_0.csv
++ cat /root/demo/tmpfile.out
+ for file in '`cat $TEMPFILE`'
+ bpplinclude demotest1 -add '\\nbuclient.com\enterprise_sales_2019\italy_sales.rtf'
+ for file in '`cat $TEMPFILE`'
+ bpplinclude demotest1 -add '\\nbuclient.com\enterprise_sales_2019\ip_cidrs.txt'
+ for file in '`cat $TEMPFILE`'
+ bpplinclude demotest1 -add '\\nbuclient.com\enterprise_sales_2019\console.sh'
+ for file in '`cat $TEMPFILE`'
+ bpplinclude demotest1 -add '\\nbuclient.com\enterprise_sales_2019\uk_dl.rtf'
+ bppllist demotest1 -U
------------------------------------------------------------------------

Policy Name:       demotest1

  Policy Type:        Standard
  Active:             yes
  Effective date:     10/15/2019 16:33:33
  Client Compress:    no
  Follow NFS Mounts: no
  Cross Mount Points: no
  Collect TIR info: no
  Block Incremental: no
  Mult. Data Streams: no
  Client Encrypt:     no
  Checkpoint:         no
  Policy Priority:    0
  Max Jobs/Policy:    Unlimited
  Disaster Recovery:  0
  Collect BMR info: no
```

```
Residence:         (specific storage unit not required)
Volume Pool:       NetBackup
Server Group:      *ANY*
Keyword:           (none specified)
Data Classification:      -
Residence is Storage Lifecycle Policy: no
Application Discovery:     no
Discovery Lifetime:     0 seconds
ASC Application and attributes: (none defined)

Granular Restore Info: no
Ignore Client Direct: no
Use Accelerator: no

Clients:     (none defined)

Include: \\nbumaster.com\enterprise_sales_2019\italy_sales.rtf
       \\nbumaster.com\enterprise_sales_2019\ip_cidrs.txt
       \\nbumaster.com\enterprise_sales_2019\console.sh
       \\nbumaster.com\enterprise_sales_2019\uk_dl.rtf

Schedule:          mysched01a
  Type:            Full Backup
  Frequency:       every 7 days
  Excluded Dates----------
    No specific exclude dates entered
    No exclude days of week entered
  Synthetic:       0
  Checksum Change Detection: 0
  PFI Recovery:      0
  Maximum MPX:       1
  Retention Level:   1 (2 weeks)
  Number Copies:     1
  Fail on Error:     0
  Residence:         stu001
  Volume Pool: (same as policy volume pool)
  Server Group: (same as specified for policy)
  Residence is Storage Lifecycle Policy: 0
```

```
  Daily Windows:
      Sunday    23:00:00 --> Sunday    24:00:00
      Monday    23:00:00 --> Monday    24:00:00
      Tuesday 23:00:00 --> Tuesday 24:00:00
      Wednesday 23:00:00 --> Wednesday 24:00:00
      Thursday 23:00:00 --> Thursday 24:00:00
      Friday    23:00:00 --> Friday    24:00:00
      Saturday 23:00:00 --> Saturday 24:00:00

+ bpplclients
Hardware       OS                    Client
-------------- ------------------------------ ----------------
Windows-x64    Windows               nbuclient.com
Linux          RedHat2.6.32          nbumaster.com
+ bpplclients demotest1 -M nbumaster.com -add nbuclient Windows-x64 MS-Windows
+ bppllist demotest1 -U
+ grep HW
  HW/OS/Client: Windows-x64 MS-Windows nbuclient
+ echo 'Done!!'
Done!!
```

Figure 12 shows what the created policy looks like in the NetBackup administration console.

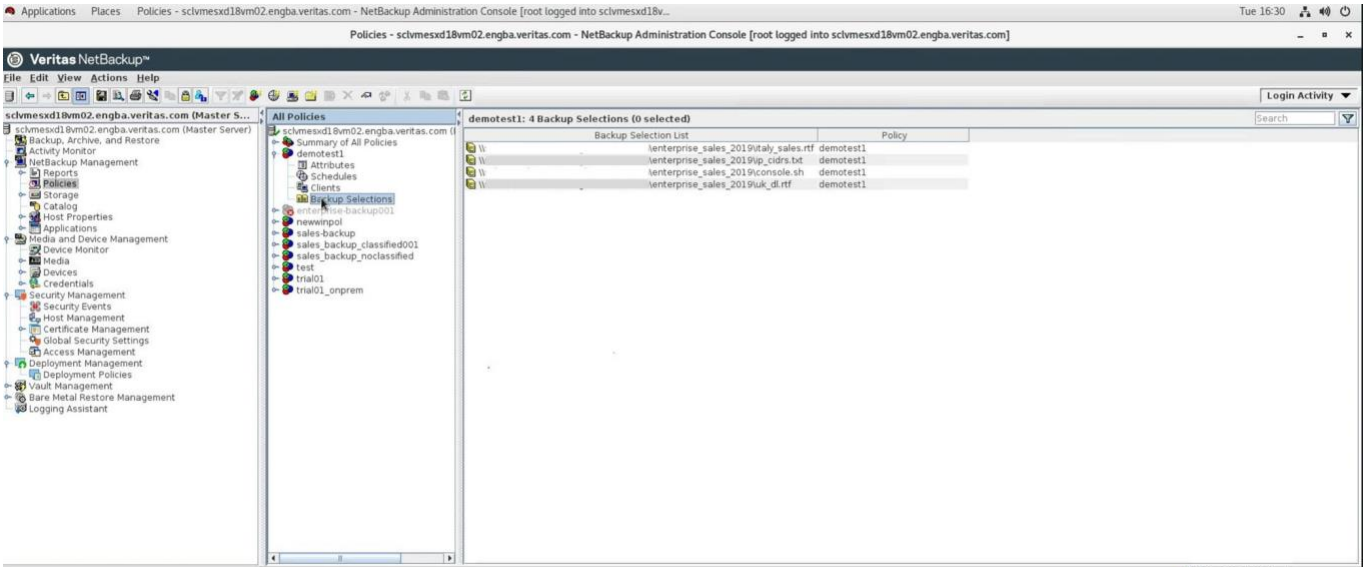


Figure 12.The created policy as it appears in the NetBackup administration console.

CONCLUSION

Cloud service providers offer varying storage tiers or classes at different levels of availability and cost to meet the various needs of customers. Information Studio is instrumental in assisting organizations to identify data they can safely move to the cloud to minimize risk and liability. Using cloud object storage with NetBackup provides a compelling option for protecting and preserving data from on-premises challenges such as failures, attacks and disasters as well as providing an off-premises long-term retention solution.

VERITAS REFERENCES

- **NetBackup**
  - Product Documentation
    - https://www.veritas.com/support/en_US/article.DOC5332
  - NetBackup Deduplication Guide
    - https://www.veritas.com/content/support/en_US/doc/25074086-136046435-0/index
  - NetBackup Cloud Administrator's Guide
    - https://www.veritas.com/content/support/en_US/doc/58500769-135186602-0/v58383369-135186602
  - Disaster Recovery
    - Veritas NetBackup in Highly Available Environments Administrator's Guide
      - https://www.veritas.com/support/en_US/doc/39129704-133515633-0
    - Disaster Recovery Procedure for CloudCatalyst
      - https://www.veritas.com/content/support/en_US/doc/25074086-127355784-0/v127468421-127355784
- **Access Appliance**
  - Product Documentation
    - https://www.veritas.com/content/support/en_US/dpp.Access.html
  - White Papers/Data Sheets
    - https://www.veritas.com/protection/access-appliance/resources
- **Information Studio**
  - Product Documentation
    - https://sort.veritas.com/documents/doc_details/INFOSTUDIO/1.1/General/Documentation/
  - White Paper
    - https://www.veritas.com/content/dam/Veritas/docs/white-papers/V0956_WP_Information-Studio.pdf

DISCLAIMER

ABOUT VERITAS

Veritas Technologies is a global leader in data protection and availability. Over 50,000 enterprises—including 87 percent of the Fortune Global 500—rely on us to abstract IT complexity and simplify data management. The Veritas Enterprise Data Services Platform automates the protection and orchestrates the recovery of data everywhere it lives, ensures 24/7 availability of business-critical applications, and provides enterprises with the insights they need to comply with evolving data regulations. With a reputation for reliability at scale and a deployment model to fit any need, Veritas Enterprise Data Services Platform supports more than 800 different data sources, over 100 different operating systems, more than 1,400 storage targets, and more than 60 different cloud platforms. Learn more at www.veritas.com. Follow us on Twitter at @veritastechllc.

Veritas World Headquarters

2625 Augustine Drive

Santa Clara, CA 95054

+1 (866) 837 4827

www.veritas.com

For specific country offices and contact numbers, please visit our website.