WHITE PAPER: ENTERPRISE SOLUTIONS

Secure Optimized Data Protection for Remote Offices

An Overview of Veritas NetBackup PureDisk™ Remote Office Edition

Wim De Wispelaere
Senior Manager, Product Management

White Paper: Enterprise Solutions

Secure Optimized Data Protection for Remote Offices White Paper

An Overview of Veritas NetBackup PureDisk™ Remote Office Edition

Contents

Data management landscape

Many organizations are struggling with the massive changes in data storage requirements that have transpired over the last decade. The almost exponential growth of business-critical data from email, e-commerce, and electronic systems shows no sign of decreasing. With relatively new data types such as voice and video now in use, enterprise storage administrators will soon have to manage petabytes of data.

According to a recent study by the School of Information Management and Systems at the University of California at Berkeley, the world produces between 1 and 2 exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. An exabyte is a billion gigabytes and printed documents of all kinds comprise only .003 percent of the total. Magnetic storage is by far the largest medium for storing information and is the most rapidly growing, with shipped hard-drive capacity doubling every year. Magnetic storage is becoming the universal medium for information storage.

The annual growth rate of corporate reference data is estimated to be 60 percent. Traditional ”weekly full, daily incremental” backup approaches are ill suited to cope with this situation: companies should work 60 percent more efficiently to prevent costs from increasing, which is not very likely to happen with the current systems. As the business world has moved to a 24-hour, 7-day-a-week working cycle, the notion of overnight “downtime” for maintenance and backup is less feasible. Symantec has developed software to help companies manage data integrity more efficiently and meet today’s standards for data protection.

The rise in the volume of data has seen a corresponding tightening of corporate governance and legal procedures surrounding the retention and availability of data. According to one large storage vendor, there are over 4,000 major regulations that apply to information-keeping worldwide. The United States has the most, with federal statutes such as the Health Insurance Portability and Accountability Act (HIPAA) that covers medical records, and the Food and Drug Administration Section 21 rules that carry heavy fines for noncompliance with data-retention rules. The most commonly quoted U.S. regulation is the Sarbanes-Oxley Act of 2002, which was brought into force after Andersen employees shredded important documents in the wake of the Enron scandal. The Securities and Exchange Commission (SEC) also has extensive rules governing data retention, with heavy fines and even jail sentences for executives in cases of noncompliance.

The U.S. National Archives and Records Administration (NARA) and the United Kingdom's Public Records Office are two examples of bodies whose entire business is focused on ensuring records are maintained correctly and effectively. In fact, most developed countries have similar governmental structures to ensure that data is archived and released within a legal framework. These guidelines, in turn, must be adhered to by other governmental agencies such as the Social Security Administration and the Internal Revenue Service, as well as banks and building societies.

In Europe, draft European Union (EU) legislation being formally ratified now will force telecom companies and Internet service providers (ISPs) to retain information on their customers' logs of phone calls or e-mail and Internet connections beyond the one- or two-month period the information is normally held for billing purposes. The period could be up to a year; this change is intended to assist police and fraud investigations.

Data and risk evolution

Conventional data management and protection techniques have not kept pace with the increasingly complex nature of today's data processing and topology. Many IT organizations still use the "weekly full, daily incremental" backup technique employed since the 1950s. In the past half century, topologies have evolved from centralized homogeneous platforms to heterogeneous networks of distributed and mobile systems with multiple storage tiers; annual data growth has increased from 20–35 percent to 80–100 percent; retention periods have increased from weeks to decades; and usage patterns have evolved from transaction to transaction and reference. Studies indicate approximately 60 to 80 percent of this growth is fuelled by reference data. Reference data describes information with access requirements measured in seconds to minutes while transaction data describes information with access requirements in milliseconds.

Traditional data-protection and management approaches fail. Ernst & Young's Fabric of Risk study determined that approximately 36 percent of the executives from the top 1,000 publicly traded companies believed their companies would cease operations due to inadequate protection, while 59 percent placed their risk as moderate to high. The increased dependence upon networked and mobile data, combined with theft and vulnerability to viruses, exacerbates this risk. New technologies such as storage area networks (SANs) aid physical storage management, but affordable data-protection and management solutions have been elusive.

Present-day issues with classic technology

The globalization of businesses and cultural changes such as home and mobile working have made the backup process more complex. Companies operating from different local offices must distribute parts of their IT infrastructure over these remote sites out of necessity. Local documents, emails, presentations, and so forth are kept on local file servers primarily to improve network performance and to allow rapid recovery in the event of data loss (Figure 1).

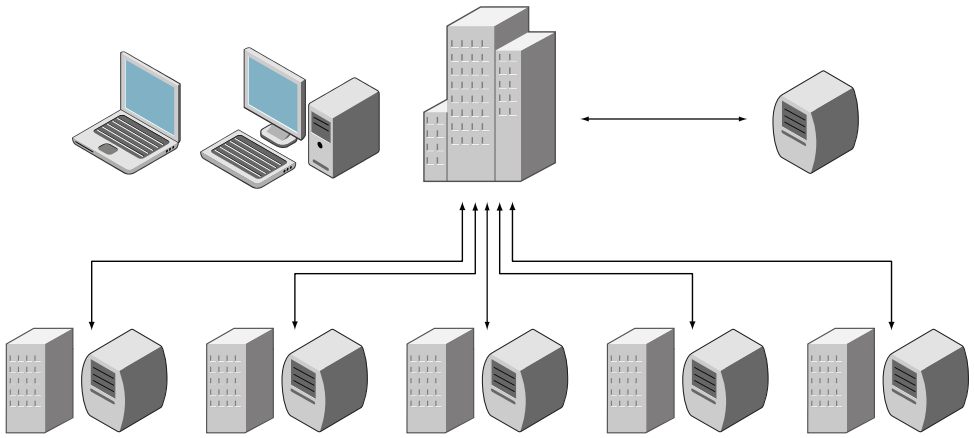

Figure 1. The volume of files on remote office servers continues to grow rapidly.

Many enterprises organize the backups of their remote sites locally, rather than sending data over a wide area network (WAN) to a central backup server. Local backup, however, isn't without its hazards:

- It can hinder central control and monitoring and hence introduces the potential for errors.
- The people performing remote backups may not follow central procedures and security policies exactly as specified. Are tapes systematically transferred to a vault and stored in a secure environment that protects them against changes in temperature and humidity? Is the backup executed every day, as stated in the guidelines?
- Backup and recovery operations can fail due to a lack of skilled resources, lack of media, or other technical problems not easily solved remotely.

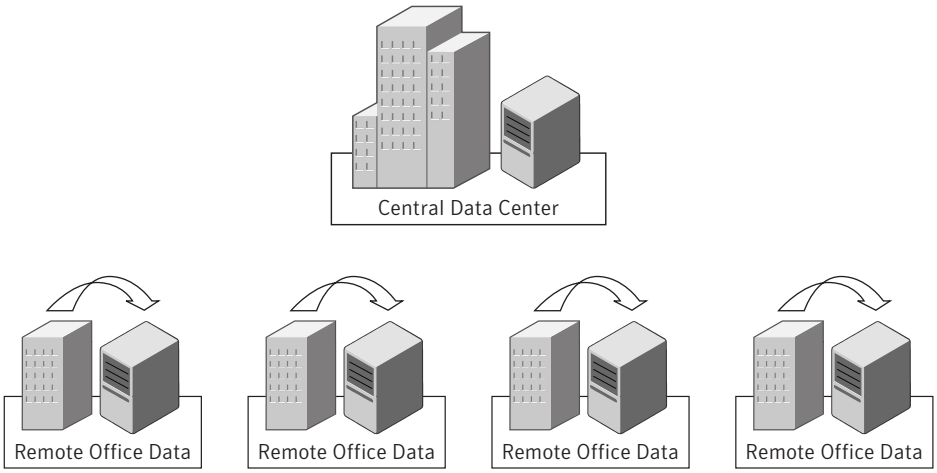

Figure 2. Remote office backup can be cumbersome and problematic.

As many studies and onsite audits have pointed out, the backup of data at a remote site is often executed by non-IT personnel. Sometimes the wrong tape is inserted and there is no one else present to check it. The response to system errors is often incorrect and is not reported to a central IT authority. Worse, when the tape loading process fails, no backup is executed. As the remote employee is often not qualified to verify whether the backup has been successful or not, no one really knows if the tapes contain the correct data.

Limitations of the current WAN solution

The number and size of files stored on local file servers are increasing dramatically. Consequently, performing a full backup of remote data over a WAN connection requires a considerable amount of bandwidth and can be prohibitively expensive. As the number of remote sites increases, this problem becomes an even bigger issue, creating a bottleneck between the remote site and the data center's backup server. As data volumes grow and working patterns edge towards 24 x 7 operation, the overnight backup window may not even be a feasible option. This leaves many sites with only one option: to organize all backups locally with backup tapes stored at the site or sent to an offsite facility.

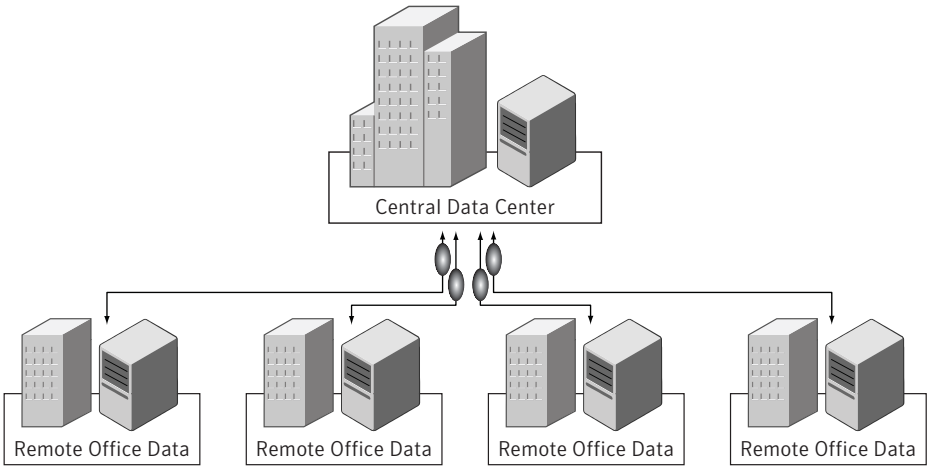

Figure 3. Corporate wide-area networks (WANs) can not handle traditional backups of remote offices.

NetBackup PureDisk Business Proposition

The massive growth in data generation and retention periods, combined with legislative requirements, requires a fundamental change in backup procedures. This is especially true with remote offices and mobile users who have often been outside of the scope of a centralized IT infrastructure.

Symantec is a pioneer in the field of content routing, a technology that provides a long-term framework to address both the rapidly growing volume of data and the wider information lifecycle issues. Veritas NetBackup PureDisk is a software solution for the protection of file data on clients anywhere on the network to any type of disk-based storage pool.

With its unique fingerprint technology called global unique file identification, NetBackup PureDisk technology distinguishes unique files from redundant copies across the enterprise. Enormous savings in storage capacity and network traffic are achieved by not transmitting and storing redundant data.

For example, a backup across three remote offices of the same two-megabyte Word file on three file servers would result in six megabytes of capacity used in conventional approaches. NetBackup PureDisk, however, stores a single copy only and consequently needs no more than two megabytes of storage capacity—a 66 percent savings. Comparable results are seen for throughput: The backup would be completed about 66 percent faster across the three remote sites.

Through file segmentation, this fingerprint technology can even be applied on small parts of files for global unique segment identification. This is typically done for large files such as .pst files. When such a file is backed up and then modified, only the modified segment(s) will be backed up.

Global unique file and segment identification is performed by lightweight PureDisk agents on the client systems. The backup data is stored centrally in a PureDisk Storage Pool. The Metabase in the PureDisk Storage Pool will store the file metadata in a scalable, distributed database. The file content segments are stored in one or multiple Content Routers in the Storage Pool. Because metadata and content data are stored separately, all source file versions can be restored, while only the globally unique file segments are stored.

In a typical system, about 13 percent of the files are modified each day and must be backed up. Using PureDisk global unique file and file segment identification, our studies show the number of actual bytes changed is 1–2 percent. For example, say a company has one terabyte of data. If it performs four incremental and one full backup each week over a period of 16 weeks, these 80 backups would require more than 16 terabytes of storage. PureDisk will store an optimized initial backup and then 16 weeks of incremental backups. After 16 weeks, PureDisk will typically need 1.5 terabytes to store all versions of all files within this retention time. With 80 terabytes of source data protected in 80 backups of one terabyte each, PureDisk has reduced the backup data volume by more than 50 times to 1.5 terabytes.

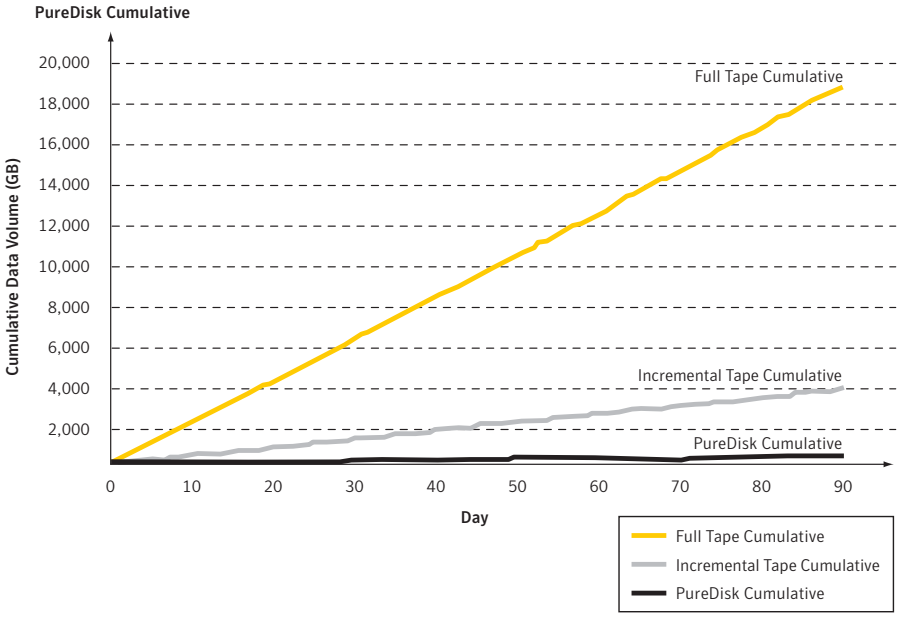

Figure 4. Growth of backup data volume using different backup methods.

PureDisk agents installed on the clients perform the global unique file and segment identification locally and only send incremental data over the network. This can reduce the backup bandwidth by a factor of 50. Figure 5 illustrates the bandwidth savings of NetBackup PureDisk versus traditional methods. Because only new and unique file segments are transmitted, present-day WANs such as Internet VPN have sufficient bandwidth for the transfer of remote data to the data center.

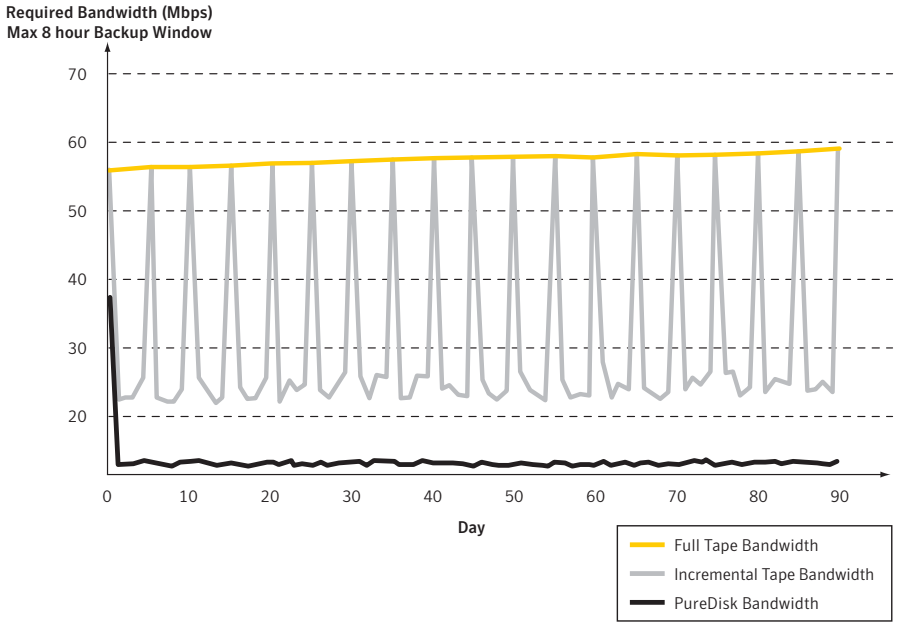

Figure 5. Comparison of bandwidth required to replicate 1 TB of data over WAN within 8 hours.

NetBackup PureDisk reduces the maintenance cost related to backing up data at remote offices. Typically, such maintenance involves system maintenance, tape management, and the shipment of media to a vault site. NetBackup PureDisk performs automated disk-based backups and can be remotely operated using the Web-based user interface. This means NetBackup PureDisk does not require anybody to operate the system or handle tapes at the remote site. Eliminating manual backup system handling will result in a higher backup-and-restore success rate.

Because PureDisk backup data is replicated automatically to the central office and/or a recovery site, there is no need to ship physical media. This results in additional cost savings.

NetBackup PureDisk performs a data reduction of 50 times on the backup data. As a result, smaller amounts of data need to be managed, which requires fewer storage operators. Because of the data reduction, backups will be completed faster. In combination with the parallel processing capability of NetBackup PureDisk, this results in a significant reduction in overall backup time.

NetBackup PureDisk implementation

A typical environment protected by NetBackup PureDisk has a PureDisk Storage Pool at each site for local backups and a central Storage Pool for replication of the local backup data. However, PureDisk technology can also protect smaller environments without the need for a local Storage Pool. Figure 5 displays both of these deployment scenarios.

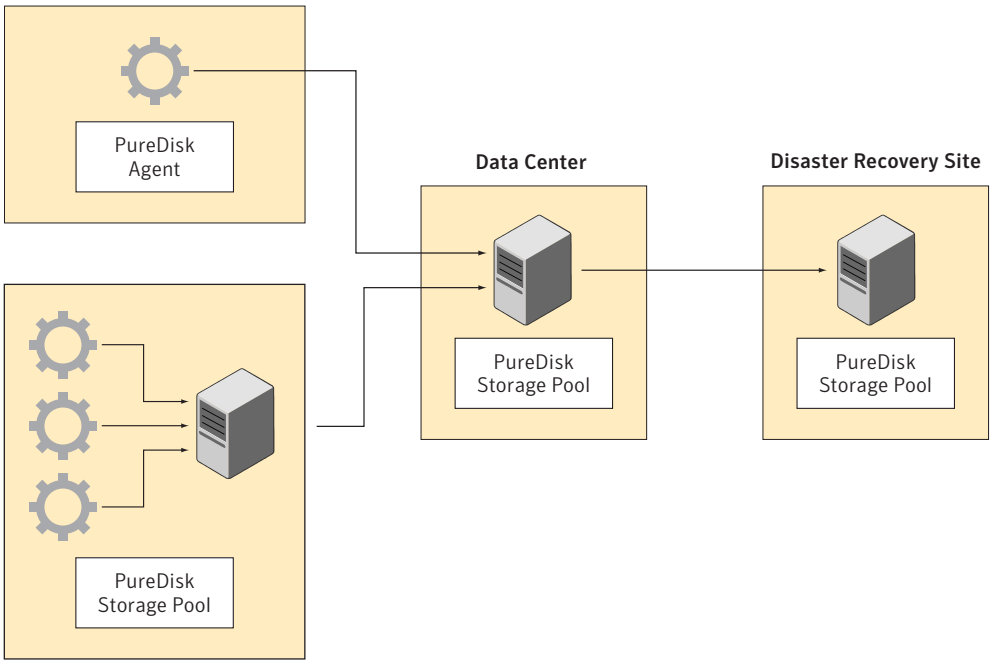

Figure 6. Illustrates how agents can backup data to local or remote storage pools and how these storage pools can efficiently replicate to other remote storage pools.

The agents on all systems send the new, unique file segments to a local PureDisk Storage Pool, which verifies the uniqueness of files across all local agents. It then replicates only the unique file segments to the main PureDisk Storage Pool. During this replication, the uniqueness of the file segments is checked across all connected agents and Storage Pools. This minimizes WAN bandwidth requirements and allows long-term scalability because of the reduced Storage Pool capacity requirements.

A PureDisk Storage Pool on the remote site optimizes the data over all the office's clients by identifying unique contents and stores backup data locally. This shortens backup-and-restore tasks, and enables synchronization with central or distributed locations independent of the local backups.

While backups are performed over the LAN during business hours, synchronization to the central data center can be performed overnight, when the WAN has less activity. This reduces WAN traffic back and forth between the data center and remote sites.

Using PureDisk replication, there are always two full remote-site backups—one local, and one on a central server—every day. As an option, backup data can be replicated again to a disaster recovery site, adding another level of data protection.

In a replication policy, data filters can be applied. Using filters, only selected files or file types will be replicated to the central Storage Pool.

Data retention in the PureDisk environment is managed by data removal policies per Storage Pool. In the remote office, a data retention policy can be defined to retain all versions of files during a short period of time, typically a few weeks. On the central Storage Pool, a data retention policy for the replicated data can be configured to keep all versions of all files for a longer period of time, typically several months. Using the PureDisk data retention features, the local PureDisk Storage Pool can be configured as a local backup cache allowing fast local restores, while the central Storage Pool can be configured as a medium-term data archive.

Smaller remote offices with smaller amounts of data or fewer clients to protect can opt not to have the local PureDisk Storage Pool at the remote site. Instead, they can back up files instantly over the WAN to the central Storage Pool. All files, however, are secured using 256–bit, key agent–based encryption.

This solution is less expensive to implement because it eliminates the need for a device at the remote site. Performing the backups over the WAN will slow down backup-and-restore tasks; these will be limited by the WAN bandwidth. This option can be used if the recovery time objective can be met with the available WAN bandwidth. This scenario is only recommended if single file restores over the WAN are required. Full data restores should be performed at the central site to a spare system or to removable media, which can then be shipped to wherever the data is needed.

PureDisk technology and architecture

Symantec uses the term Global Single Instance Storage to represent the combination of Global Unique File and Segment Identification.

Global unique file identification

PureDisk employs a distinctive fingerprint technology to distinguish unique files from redundant copies. The fingerprint is derived from the total binary contents of the file. The result is that files with the same content will have the same fingerprint, even when the files have different names, locations, attributes, creation or modification dates, and security. Only one copy of a file with a

certain fingerprint will effectively be sent to the PureDisk Storage Pool. Other copies of the same file will be identified because they will generate the same fingerprint. Even when the file name and path of the copied file is different or when the files are stored on geographically distant systems, NetBackup PureDisk will identify these files to be exact copies based on their identical fingerprints. By using globally unique fingerprints, the disk space required for the PureDisk Storage Pool grows at a much slower rate than by using traditional backup methods.

Global unique segment identification

NetBackup PureDisk also divides larger files into smaller segments with a configurable segment size. By cutting a large file into small chunks of data, the system will only back up the parts that contain new or modified data, while the majority of the file remains unchanged and doesn't need to be backed up again.

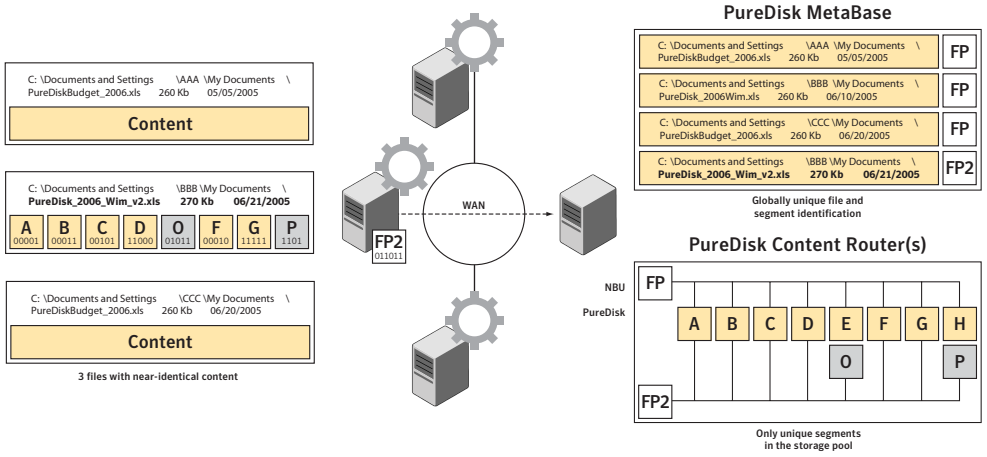

Figure 7. Global Unique File and Segment Identification.

Separated storage of metadata and content data

Agents send the backup data to a PureDisk Storage Pool. In a PureDisk Storage Pool the file properties are stored in a Metabase, and the file contents are stored in Content Router(s).

For all files backed up, including file versions and deletions, the file metadata is stored in the Metabase. The Metabase is a scalable, fully searchable, distributed database for metadata. The Metabase contains the file metadata such as name, path attributes, and security settings, together with the file fingerprint.

The unique file contents are stored in one or more Content Router(s). The first characters of the segment fingerprint determine the position where the file segment is stored in the Content Router(s) according to the content routing table.

With all metadata of all file versions in the Metabase, you can restore any version of any file to any client, while in the Content Router(s) only the necessary unique file segments have been stored.

Metabase

The PureDisk Metabase stores all metadata of files and previous versions, together with their fingerprints. The front end of the PureDisk Metabase is called the PureDisk Server. The PureDisk Metabase Server delivers a fully searchable view across all the stored metadata. The actual metadata is stored in one or more PureDisk Metabase Engines (see Figure 8). The Metabase Engines are relational database systems. Through its two-layer architecture, the Metabase can be scaled in size and performance by simply adding Metabase engines.

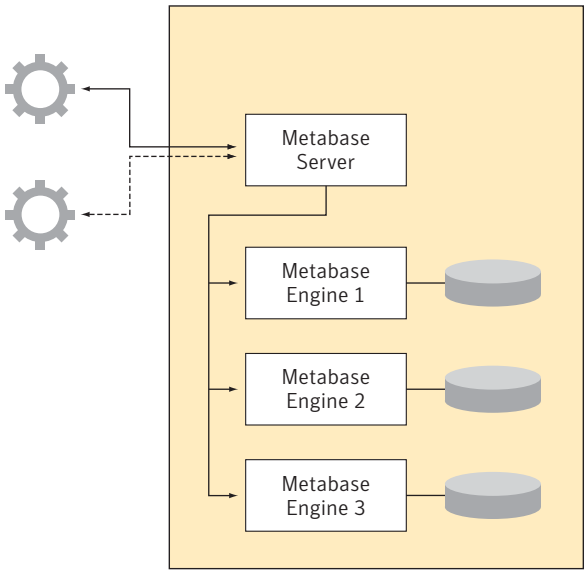

Figure 8. NetBackup PureDisk uses a 2-tier architecture for storage of metadata.

Users can query the Metabase to locate a particular file or set of files. The Metabase search tools allow the user to specify many different file-property and time-related search criteria. Apart from its obvious assignment as a keeper of backup metadata, the Metabase is also a valuable tool

for data assessment. For example, a user could query the Metabase to identify all data older than three months across all of its remote file servers, or all data that was not accessed in the last month. Complex questions can be answered by combining queries and post-processing the results. For example:

- How much storage space do I currently need to store all my documents or all my presentations?
- How much duplicate information is currently stored?
- How many versions of this document do we have?
- Which systems use this infected dll?

Content Router

A Content Router is responsible for storing and restoring the file segment data from and to the PureDisk Agents. A Content Router can also redirect file content data to another Content Router.

Each Content Router is connected to a storage device of any type (direct attached storage [DAS], network attached storage [NAS], SAN). When receiving content-enabled data for which the current Content Router is the final destination, it will effectively store this data on its connected storage device.

Each Content Router is responsible for a specific subrange of the full PureDisk content address range. The address of a data object is derived from the unique fingerprint of this object. The distribution of the content addresses across the available Content Router(s) is determined in the content routing table.

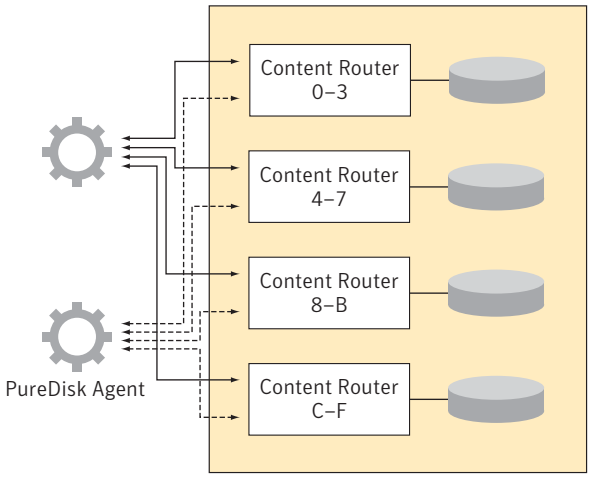

Figure 9. Content Routers store and restore the unique file segment data through the PureDisk Agents.

The number of Content Routers in a PureDisk environment depends on the amount of data to be dealt with and the data traffic. Scalability is achieved by simply adding more Content Routers (with attached storage) when the need arises without having to interrupt or modify the backup or restore policies. This provides online scaling beyond petabytes of data within a single Storage Pool. The expansion of the total volume of storage can happen while the PureDisk system remains online. With the addition of new Content Routers with storage, the content routing tables will be updated to reflect the new configuration. From that moment, the Content Routers will redistribute the existing content data according to the new storage organization. This happens in the background while the system remains available for backups and restores.

Web-based user interface

All PureDisk functionality is managed through an easy-to-use Web interface. This interface is accessible from any Java-enabled Web browser without having to install any additional software. The Web interface is accessed through an SSL secured connection, making it safe to manage the PureDisk Storage Pool from a remote location.

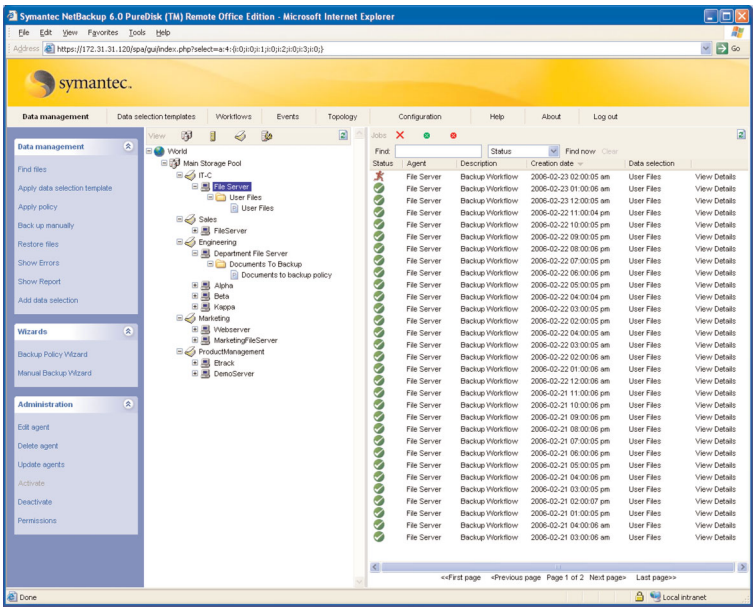

Figure 10. The NetBackup PureDisk Web-based interface.

An administrator defines the visible scope and the functional level for each user that logs on to the Web interface. The scope can be limited per client, group of clients, or location. The functional level can allow one user to only restore files while another user can be given the right to manage the configuration of a Storage Pool.

PureDisk agent-based restores are scheduled or initialized via the Web interface. Operators or users with the required access rights can manage agent-based restores from any location. Because PureDisk is disk-based, a restore in a remote location can be initiated centrally with no manual intervention onsite.

Web restore interface

In addition to agent-based backup-and-restore functionality, the Web-based search tool allows the user to find old files or file versions quickly and download these files via HTTP to any location without the need for a PureDisk agent.

Virtual file system interface

PureDisk also can be accessed through the PureDisk virtual file system interface. The PureDisk virtual file system interface is a CIFS interface that represents Metabase information in network shares. These network shares can be accessed from any CIFS-capable client.

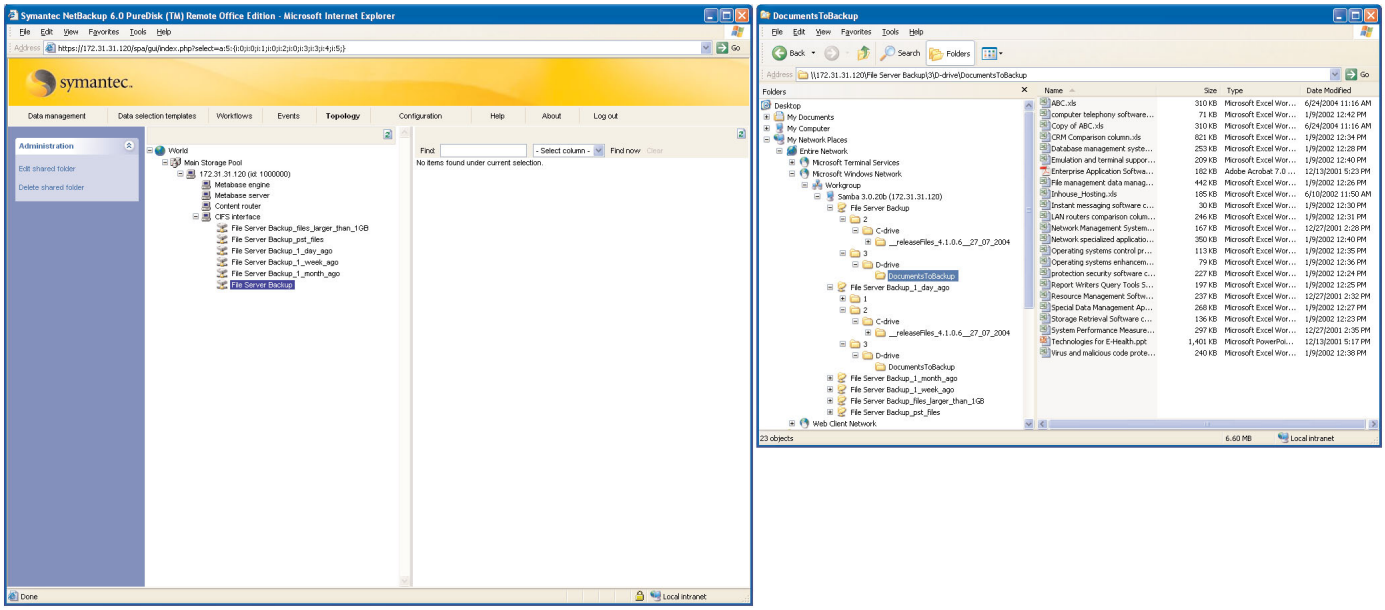

Figure 11. The NetBackup PureDisk virtual file system interface.

The information presented in the CIFS mount point is filtered metadata. Filters can be defined based on client information, file type, path inclusions and exclusions, and source system presence time. For example, an operator can define mount points in the virtual file system that represent client data as it was present during the last backup, one week ago, and one month ago. If a file is copied from such a virtual mount point, the virtual file system will decompress, de-encrypt, and reassemble the file on the fly based on the data from the Metabase and the file segments from the Content Router(s).

Parallel processing

Backup processes are distributed to individual clients at both central and remote sites throughout the entire enterprise network. Scalability is guaranteed because large portions of the backup process for each client is performed locally on the system by an agent. This allows backups of hundreds of clients to run simultaneously.

Each client will only send the globally unique file segments to its Storage Pool. Even when hundreds of agents are performing backups, the volume of backup data on the network remains limited. Also, the Storage Pool needs to accept only a minimal volume of data. Thus, the parallel processing of hundreds of agents will not flood the network, nor will it stress the Storage Pool servers. The result is a highly scalable backup environment that is virtually insensitive to the number of clients and remote sites in your business IT environment.

Backup data encryption

The PureDisk agent optionally encrypts every file and file segment that it sends to the Storage Pool. Encryption of the data before sending it over the WAN to the Storage Pool ensures that the data is secure at all times—the data is unreadable and unusable by any hacker who may attempt to tap the data stream.

All encrypted file segments that are received by the Storage Pool are stored to the Content Router disks in their encrypted form. No description is performed, resulting in end-to-end data security. The disk architecture eliminates the risk of accidental loss of tapes, and unauthorized personnel cannot gain access to the data stored in the disk-based Storage Pool.

Policy-driven retention and data removal

By default, PureDisk will retain all versions of all protected and migrated data in the Storage Pool. Storage administrators can define policies that will remove obsolete data from the storage system. These policies can be applied on any selection of data sources. Removal policies can use time (older than), version (last n versions) and file attributes (extension, size) to filter the files to be removed.

Storage pool replication and disaster protection

Because of its unique content routing architecture, PureDisk can replicate its storage to multiple local and/or remote storage devices. Content optimization will keep the required bandwidth and the replication time to a minimum. By replicating to Content Routers and storage devices in different locations, an additional layer of protection is added to safeguard against potential disaster at the main site. The complete PureDisk architecture can be rebuilt and all the machines under backup control can be restored from the data at the remote site.

Replication to the remote site can be performed in asynchronous mode. This guarantees that those backups are performed in the LAN for optimal backup performance and a short backup window, while the offsite replication can take place at a slower rate.

Conclusion

- The fact is that more data will be generated and held on disparate computer systems. Organizations of all sizes need to ensure that their data is properly secured with the ability to store, access, recover, and ultimately remove it. Pressure from legislation, combined with pressing business needs, require that data protection policies of proven benefit be in place.
- Many of the existing data-protection strategies are based on obsolete business practices such as a 10-hour working day and overnight backup windows. The modern world of 24-hour-a-day industry and electronic transactions has forced an increased reliance on data systems.
- NetBackup PureDisk combines breakthrough technology and modern data protection concepts to put the brakes on the exponential increase of backup data and bandwidth. It uses fingerprint technology to distinguish unique files and file segments from redundant ones and incorporates all systems in the data management process, even those on remote sites.
- Many organizations are in the process of deploying data protection solutions to solve short-term problems. Many of these solutions revolve around one vendor or hardware platform. PureDisk offers a data protection infrastructure that provides a long-term solution with a predictable cost model—a solution with flexibility that embraces any hardware vendor or storage technology.
- You can reduce remote-office backup costs using the PureDisk policy-driven disk-to-disk backup technology that automates remote backup-and-restore tasks. Backup data is centralized using existing network connections, so manual media handling and shipments are no longer required.
- NetBackup PureDisk is a secure backup solution that only transmits and stores data after it has been encrypted. PureDisk components and the PureDisk interfaces use SSL-secured connections to communicate.
- The NetBackup PureDisk solution is able to adapt to emerging technology, legislation, and business processes. Content routing and fingerprint technology fundamentally changes the way data is handled and enables the kind of information lifecycle management that is needed now and in the foreseeable future.

About Symantec

Symantec is the world leader in providing solutions to help individuals and enterprises assure the security, availability, and integrity of their information. Headquartered in Cupertino, Calif., Symantec has operations in more than 40 countries. More information is available at www.symantec.com.

For specific country offices and contact numbers, please visit our Web site. For product information in the U.S., call toll-free 1 (800) 745 6054.

Symantec Corporation
World Headquarters
20330 Stevens Creek Boulevard
Cupertino, CA 95014 USA
1 (408) 517 8000
1 (800) 721 3934
www.symantec.com

02/06 10526109