SPECTRA

DEEP STORAGE EXPERTS

Archiving to a Private Cloud to Affordably Keep Data Forever

Use simple archive workflows from a single vendor to reduce complexity and cost

8/31/2015

CONTENTS

Abstract

Organizations in many industries are dealing with explosive data growth. In these types of organizations, typically their data is their business. Often these organizations must have an effective way to keep this data for years or essentially forever. The most affordable option is fast, reliable digital tape. Use of a private cloud in front of a tape library to move this data makes it easy with a complete, single-vendor solution. Spectra Logic provides tools that can create a complete deep storage archive solution for many basic workflows. These solutions reduce the cost of storage of this data from dollars (USD) per gigabyte to pennies per gigabyte. Public cloud is not an option when an organization has petabytes of data that must be stored for an extended period of time.

Introduction

Many organizations are dealing with exponential data growth. Industries such as genomics, media and entertainment, research, and national governments are often dealing with generating many terabytes of data per day and want to keep all that data for extended periods of time. For these organizations, their data is their business, and they will continue to repurpose and/or re-monetize the data over many years. Additionally, regulations often require retaining the data for 100 years[1] or "the life of the Republic"[2]. Initially this data is generated on primary disk storage, which provides high performance but can cost on the order of dollars per gigabyte of storage. These organizations cannot afford to keep this data on primary disk storage for any length of time. We will use genomics research as a typical example of an organization with exponential data growth. In genomics each experiment or sequencing analysis can occur multiple times daily and generate terabytes of data. How do genomics researchers affordably store this data while still making it easily accessible?

Traditional Workflow

In a typical but undesirable workflow for these genomics researchers, because the data is being stored on primary disk, the researchers would identify the low-value data and delete it soon after creation. For the remaining higher value data, the researchers might choose to utilize a disk archive system, which is roughly half the cost per gigabyte of a primary disk storage system, on which to move some of the data. But disk archive systems are still on the order of dollars per gigabyte. Even with the disk archive system, they would eventually decide that they cannot afford to keep experimental data for more than a few years before they have to delete it. They are working under the mistaken belief that they cannot afford to keep all the data.

[1] *European Document Retention Guide*, Iron Mountain, http://www.debrauw.com/wp-content/uploads/2015/01/EU-Retention-Guide-2014.pdf

[2] http://www.digitalpreservation.gov/meetings/documents/storage13/KenWood-LoC-DSADevelopmentsInMedia2013.pdf

New Workflow - Basic

With a private cloud system, such as Spectra Logic’s BlackPearl with a Spectra tape library, data can be kept forever at pennies per gigabyte. In this case, the researchers would keep all their research data, and archive it to this private cloud as they need to free up space in their primary disk. With a simple drag and drop interface like Spectra’s free, open source Deep Storage Browser, the researchers could easily archive that data themselves. No additional storage or backup software is required.

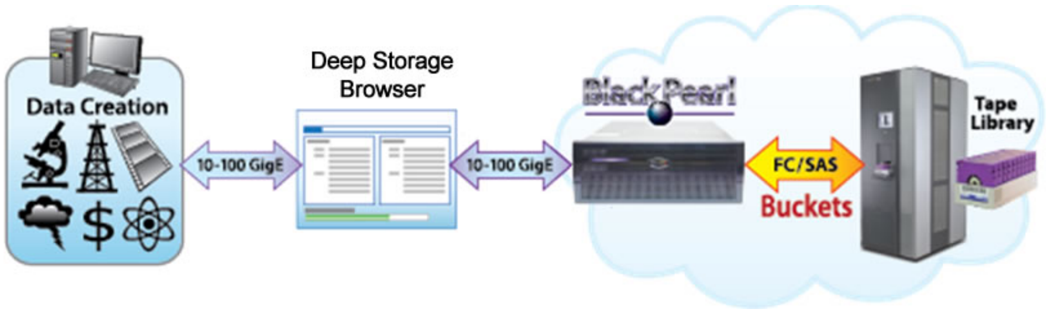
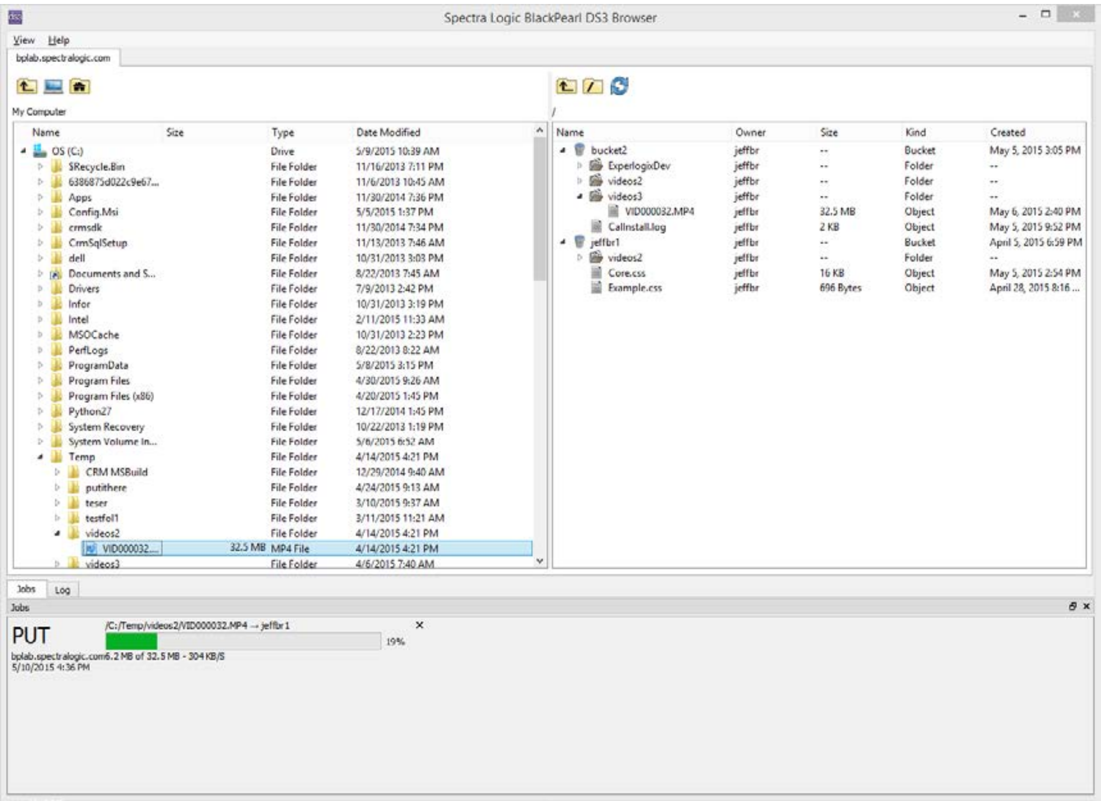


Figure 1. Archiving data with the Deep Storage Browser

The researchers would first aggregate the data into TAR or ZIP files to create a file with a minimum size of 40MB, as private cloud object storage technology provides the best throughput with this minimum files size. The researchers could then use the Deep Storage Browser to drag and drop the aggregated files into the low cost deep storage. BlackPearl can ingest data and move it to tape at a rate of about 800 megabytes per second (MB/s), which means that a 1 terabyte file can be uploaded in less than 22 minutes. The files could then be deleted from primary storage. Data durability is ensured by BlackPearl's checksum capabilities and the Spectra tape library Data Integrity Verification features. The files could be retrieved later as needed using the same Deep Storage Browser. The Deep Storage Browser includes a search feature by file name for quickly finding the right file to retrieve.



My Computer

Name	Size	Type	Date Modified
OS (C:)		Drive	5/9/2015 10:39 AM
$Recycle.Bin		File Folder	11/16/2013 7:11 PM
6386875d022c9e67...		File Folder	11/6/2013 10:45 AM
Apps		File Folder	11/30/2014 7:36 PM
Config.Msi		File Folder	5/5/2015 1:37 PM
crmsdk		File Folder	11/30/2014 7:34 PM
CrmSqlSetup		File Folder	11/13/2013 7:46 AM
dell		File Folder	10/31/2013 3:03 PM
Documents and S...		File Folder	8/22/2013 7:45 AM
Drivers		File Folder	7/9/2013 2:42 PM
Infor		File Folder	10/31/2013 3:19 PM
Intel		File Folder	2/11/2015 11:33 AM
MSOCache		File Folder	10/31/2013 2:23 PM
PerfLogs		File Folder	8/22/2013 8:22 AM
ProgramData		File Folder	5/8/2015 3:15 PM
Program Files		File Folder	4/30/2015 9:26 AM
Program Files (x86)		File Folder	4/20/2015 1:45 PM
Python27		File Folder	12/17/2014 1:45 PM
System Recovery		File Folder	10/22/2013 1:19 PM
System Volume In...		File Folder	5/6/2015 6:52 AM
Temp		File Folder	4/14/2015 4:21 PM
CRM MSBuild		File Folder	12/29/2014 9:40 AM
putithere		File Folder	4/24/2015 9:13 AM
teser		File Folder	3/10/2015 9:37 AM
testfol1		File Folder	3/11/2015 11:21 AM
videos2		File Folder	4/14/2015 4:21 PM
VID000032....	32.5 MB	MP4 File	4/14/2015 4:21 PM
videos3		File Folder	4/6/2015 7:40 AM

/

Name	Owner	Size	Kind	Created
bucket2	jeffbr	--	Bucket	May 5, 2015 3:05 PM
ExperlogixDev	jeffbr	--	Folder	--
videos2	jeffbr	--	Folder	--
videos3	jeffbr	--	Folder	--
VID000032.MP4	jeffbr	32.5 MB	Object	May 6, 2015 2:40 PM
CaInstall.log	jeffbr	2 KB	Object	May 5, 2015 9:52 PM
jeffbr1	jeffbr	--	Bucket	April 5, 2015 6:59 PM
videos2	jeffbr	--	Folder	--
Core.css	jeffbr	16 KB	Object	May 5, 2015 2:54 PM
Example.css	jeffbr	696 Bytes	Object	April 28, 2015 8:16 ...

PUT

/C:/Temp/videos2/VID000032.MP4 → jeffbr1

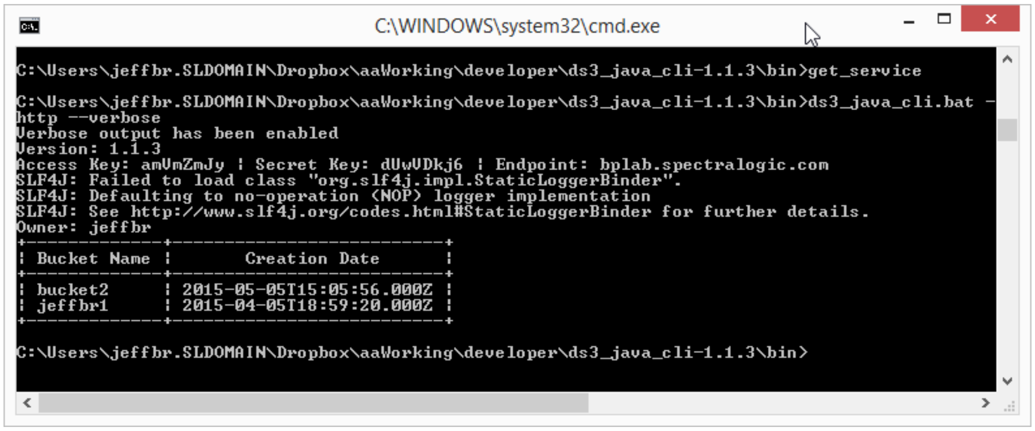
19%

bplab.spectralogic.com6.2 MB of 32.5 MB - 304 KB/S
5/10/2015 4:36 PM

Figure 2. The Deep Storage Browser

New Workflow - Intermediate

In a more sophisticated workflow, researchers could create automated scripts to move the files to their private cloud on some periodic basis. An automated script could first aggregate the data files into a TAR or ZIP file using an open-source tool such a 7Zip (7zip.og). The script could then use a tool such as Spectra's free, open source Java Command Line Interface (Java CLI) to automatically move the aggregated files to the BlackPearl private cloud. The script could be set to automatically run daily or on some periodic basis using a Cron job (UNIX) or Task Scheduler (Windows). Again, this workflow provides a complete, simple solution from a single vendor, requiring no additional software.



```
C:\Users\jeffbr.SLDOMAIN\Dropbox\aaWorking\developer\ds3_java_cli-1.1.3\bin>get_service

C:\Users\jeffbr.SLDOMAIN\Dropbox\aaWorking\developer\ds3_java_cli-1.1.3\bin>ds3_java_cli.bat -
http --verbose
Verbose output has been enabled
Version: 1.1.3
Access Key: amVmZmJy ¦ Secret Key: dUwVDkj6 ¦ Endpoint: bplab.spectralogic.com
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
Owner: jeffbr
+-------------+--------------------------+
¦ Bucket Name ¦       Creation Date      ¦
+-------------+--------------------------+
¦ bucket2     ¦ 2015-05-05T15:05:56.000Z ¦
¦ jeffbr1     ¦ 2015-04-05T18:59:20.000Z ¦
+-------------+--------------------------+

C:\Users\jeffbr.SLDOMAIN\Dropbox\aaWorking\developer\ds3_java_cli-1.1.3\bin>
```

Figure 3. The Java Command Line Interface

New Workflow - Advanced

In an advanced workflow, the researchers could create their own application to create a workflow that works best for their organization. Spectra's free, open-source Software Development Kits (SDK) can be used to create such a client application that would move data to the BlackPearl private cloud.

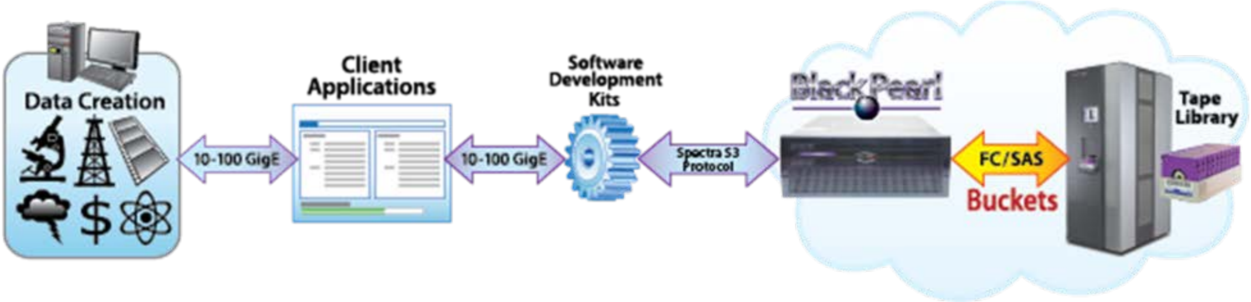
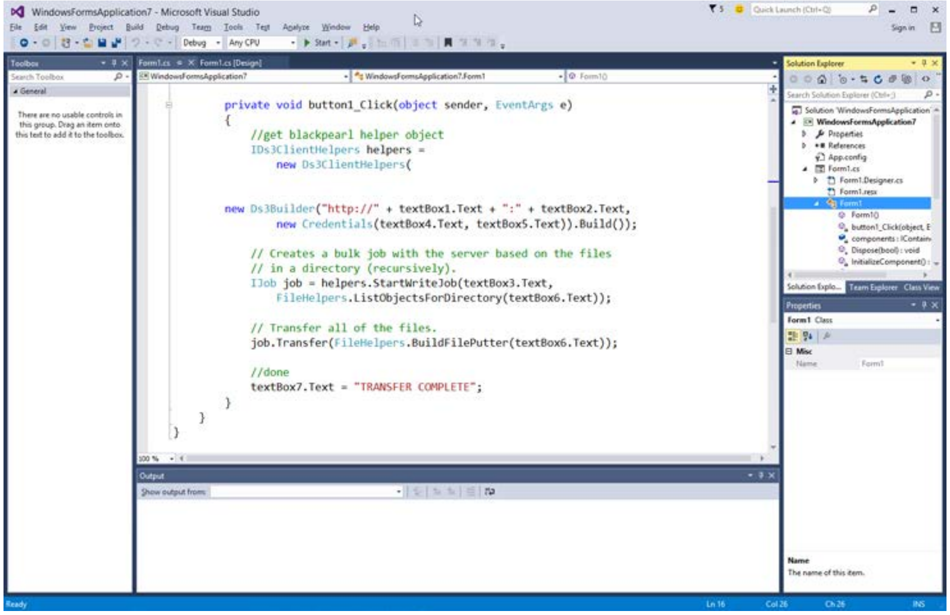


Figure 4. Archiving workflow using Spectra Software Development Kits

These SDKs are available in Java, C#/.NET, C, and Python, and allow easily movements of data to and from BlackPearl. The SDKs provide libraries that can easily be referenced in the chosen programming language. As shown in Figure 4, with just a few lines of code an entire file system directory can be easily moved to BlackPearl.


WindowsFormsApplication7 - Microsoft Visual Studio
File Edit View Project Build Debug Team Tools Test Analyze Window Help
Debug
Any CPU
Start
Toolbox
General
There are no usable controls in this group. Drag an item onto this text to add it to the toolbox.
Form1.cs
Form1.cs [Design]
Solution Explorer
Properties
Output
Ready

```
        private void button1_Click(object sender, EventArgs e)
        {
            //get blackpearl helper object
            IDs3ClientHelpers helpers =
                new Ds3ClientHelpers(


        new Ds3Builder("http://" + textBox1.Text + ":" + textBox2.Text,
                new Credentials(textBox4.Text, textBox5.Text)).Build());

            // Creates a bulk job with the server based on the files
            // in a directory (recursively).
            IJob job = helpers.StartWriteJob(textBox3.Text,
                FileHelpers.ListObjectsForDirectory(textBox6.Text));

            // Transfer all of the files.
            job.Transfer(FileHelpers.BuildFilePutter(textBox6.Text));

            //done
            textBox7.Text = "TRANSFER COMPLETE";
        }
    }
}
```

Figure 5. Using Visual Studio and the .NET SDK to Create a Simple Spectra S3 Client

Economics

By implementing a system such as the BlackPearl private cloud, the aforementioned genomics researchers' data costs can drop from dollars to pennies per gigabyte. For a genomics research organization generating 20TB of data a day (7PB/year), this can be a cost savings of millions of dollars per year.

Let's assume a primary storage system with a typical cost of $2-3 USD per gigabyte[3]. If the research organization attempted to keep all this data on primary disk storage, the purchase cost for the primary storage would be approximately $15-22 million USD. Or the research organization, faced with this staggering cost, would likely delete the data. If the data were deleted, it cannot be repurposed or re-monetized in the future.

The purchase price of a Spectra Logic T950 tape library with a BlackPearl private cloud gateway for 7 petabytes of storage is about $0.10 USD per gigabyte, or $750,000 USD (entry level systems start at $35,000 USD). This cost provides a complete solution including hardware, tape media, installation, and 3 years of support. This cost provides not only a significant cost savings, but also allows the researchers to keep all of their data rather than being forced to delete it.

And what about the public cloud? Should the public cloud be considered for storage of this data? Services like Amazon Glacier and Google Nearline offer public cloud services for as low as $0.01 USD per gigabyte per month. But for an organization generating terabytes of data per day and wanting to store it forever, there are two main problems with the public cloud:

[3] *The ROI of Primary Storage Deduplication*, Storage Switzerland, http://www.storage-switzerland.com/Articles/Entries/2012/12/3_The_ROI_of_Primary_Storage_Deduplication.html

- The per month cost of storage will add up to $1.20 USD per gigabyte over 10 years, a staggering cost more in line with primary storage.
- The internet connection bandwidth speed between the genomics organization and the public cloud would be inadequate to move 20 terabytes of data per day.

For more information about the economics of public versus private cloud, see our white paper Current and Future Economics of Deep Storage: Public vs. Private Cloud.

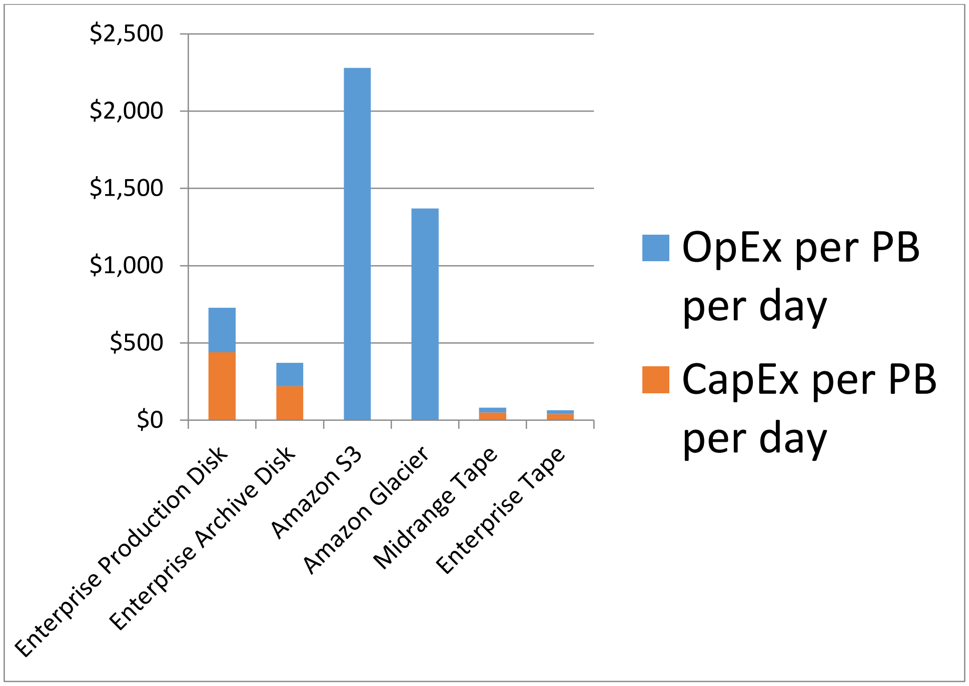


Figure 6. Storage costs per petabyte per day for storing 50PB over 5 years for various storage types

Conclusion

For organizations that generate large amounts of data, where their data is their business, they must find an affordable, durable storage solution. Primary disk storage systems are too expensive to be a long-term solution. A tape library with a private cloud interface, such as Spectra Logic's BlackPearl, provides this solution. This low-cost private cloud allows researchers and other organizations to keep this data forever, and they can keep data that may have previously considered too expensive to keep. The clients and SDKs provided by Spectra Logic provide a complete archive solution from a single vendor for many workflows, requiring no additional software.

Deep Storage Experts

Spectra Logic develops deep storage solutions that solve the problem of long term storage for business and technology professionals dealing with exponential data growth.

Dedicated solely to storage innovation for more than 35 years, Spectra Logic's uncompromising product and customer focus is proven by the largest information users in multiple vertical markets globally.

Spectra enables affordable, multi-decade data storage and access by creating new methods of managing information in all forms of deep storage—including archive, backup, cold storage, cloud, and private cloud.

For more information, please visit http://www.spectralogic.com.

SPECTRA.

DEEP STORAGE EXPERTS