

20 Years in the Making, the Future of Data Management Has Arrived

Mention data management to almost any seasoned IT professional and they will almost immediately greet the term with skepticism. While organizations have found they can manage their data within certain limits, when they remove those boundaries and attempt to do so at scale, those initiatives have historically fallen far short if not outright failed. It is time for that perception to change. 20 years in the making, Commvault [Activate](#) puts organizations in a position to finally manage their data at scale.

Those who work in IT are loath to say any feat in technology is impossible. If one looks at the capabilities of any handheld device, one can understand why they have this belief. People can pinpoint exactly where they are almost anywhere in the world to within a few feet. They can take videos, pictures, check the status of their infrastructure, text, ... you name it, handheld devices can do it.

By way of example, as I write this, I was present to watch [YY Lee](#), SVP and [Chief Strategy Officer](#) of [Anaplan](#), onstage at [Commvault GO](#). She explained how systems using artificial intelligence (AI) were able within a very short time, sometimes days, became experts at playing games such as Texas Hold'em and beat the best players in the world at them.

Despite advances such as these in technology, data management continues to bedevil large and small organizations alike. Sure, organizations may have some level of data management in place for certain applications (think email, file servers, or databases,) but when it comes to identifying and leveraging a tool to deploy data management across an enterprise at scale,

that tool has, to date, eluded organizations. This often includes the technology firms that are responsible for producing so much of the hardware that stores this data and software that produces it.

The end for this vexing enterprise challenge finally came into view with Commvault's [announcement](#) of Activate. What makes Activate different from other products that promise to provide data management at scale is that Commvault began development on this product 20 years ago in 1998.

During that time, Commvault became proficient in:

- Archiving
- Backup
- Replication
- Snapshots
- Indexing data
- Supporting multiple different operating systems and file systems
- Gathering and managing metadata

Perhaps most importantly, it established relationships and gained a foothold in enterprise organizations around the globe. This alone is what differentiates it from almost every other provider of data management software. Commvault has 20+ years of visibility into the behavior and requirements of protecting, moving, and migrating data in enterprise organizations. This insight becomes invaluable when viewed in the context of enterprise data management which has been Commvault's end game since its inception.

Activate builds on Commvault's 20 years of product development with Activate's main differentiator being its ability to stand alone apart from other Commvault software. In other words, companies do not first have to deploy Commvault's Complete Backup and Recovery or any of its other software to utilize Activate.

They can deploy Activate regardless of whatever other backup, replication, snapshot, etc. software product you may have. But because Activate draws from the same code base as the rest of Commvault's software, companies can deploy it with a great deal of confidence because of the stability of Commvault's existing code base.

Once deployed, Activate scans and indexes the data across the company's environment which can include its archives, backups, file servers, and/or data stored in the cloud. Once indexed, companies can do an assessment of the data in their environment in anticipation of taking next steps such as eDiscovery preparation, remediate data privacy risks, and index and analyze data based upon your own criteria.

Today more so than ever companies recognize they need to manage their data across the entirety of their enterprise. Delivering on this requirement requires a tool appropriately equipped and sufficiently mature to meet enterprise requirements. Commvault Activate answers this call as a software product that has been 20 years in the making to provide enterprises with the foundation they need to manage their data going forward.

**HP StoreOnce Deduplicating
Backup Appliances Put
Organizations on Path to
Ending Big Data Backup**

Headaches

During the recent HP Deep Dive Analyst Event in its Fremont, CA, offices, HP shared some notable insights into the percentage of backup jobs that complete successfully (*and unsuccessfully*) within end-user organizations. Among its observations using the anonymized data gathered from hundreds of backup assessments at end-user organizations of all sizes, HP found that over 60% of them had backup job success rates of 98% or lower, with 12% of organizations showing backup success rates of lower than 90%. Yet what is more noteworthy is through HP's use of Big Data analytics, it has identified large backups (*those that take more than 12 hours to complete*) as being the primary contributor to the backup headaches that organizations still experience.

About once every nine (9) months (*give-or-take*) HP invites storage analysts to either its Andover, MA, or Fremont, CA, offices to have a series of in-depth discussion about its portfolio of products in its Storage division. During these 2-day events, the product managers from the various groups (3PAR StoreServ, [StoreOnce Backup](#), StoreAll Archive, StoreVirtual, etc.) are given time to present to the analysts in attendance. It is during these times that candid and frank discussions ensue where each HP product is examined in-depth with the HP product managers providing context as to why they made the product design decisions that they have.

One of the more enlightening pieces of information to come out of these sessions was the amount of data that HP has collected from organizations into which its StoreOnce appliances are being considered for deployment. To date, HP has assessed environments with more than **half an exabyte** of backup data with the vast majority of backup data analyzed comprised of file system backups, either performed directly or thru NDMP.

This amount of data gives HP a rather unique perspective on

backup successes and failures. For instance, HP shared that of the approximately 4.5 million backup jobs for which it has collected data, 94.7% of them have completed successfully.

HP also revealed that organizations in particular struggle with long-running backups. Over 50% of the assessed environments had backup windows of 24 hours or more. Of these, 30% of the organizations that it had assessed had at least one backup that ran in excess of 192 or more hours – or 8 days or more. Further, the data indicates a correlation between file system backups and long backup windows.

Granted, these statistics from HP are by no means “*official*” and subject to some interpretation. However they possibly provide some of the first, large scale empirical evidence that for the vast majority of organizations that data growth goes hand-in-hand with elongated backup windows and is a major contributor if not the primary source of why backups still fail today.

Organizations moving to StoreOnce appliances, which provide high levels of performance in conjunction with source-side deduplication, are addressing this common organizational pain point as they both shorten backup windows and increase the probability that backups complete successfully. Further, using HP’s StoreOnce [Recovery Manager Central](#) solution, organizations may perform virtual machine and file system backups based on block level changes as backup data flows from HP 3PAR StoreServ to StoreOnce. This combination of solutions provides the keys that organizations need to solve backup in their environments as many organizations using the HP StoreOnce deduplicating backup appliances have already discovered.

Software Fueling Dell's Transformation to Solutions Provider; Interview with Dell Software's General Manager, Data Protection, Brett Roscoe, Part VI

Think "Dell" and you may think "PCs," "servers," or, even more broadly, "computer hardware." If so, you are missing out on one of the biggest transformations going on among technology providers today as, over the last 5+ years, Dell has acquired multiple software companies and is using that intellectual property (IP) to drive its internal turnaround. In this sixth installment of my interview series with Brett Roscoe, General Manager, Data Protection for Dell Software, we discuss how these software acquisitions are fueling Dell's transformation from a hardware provider into becoming a solutions provider.

Jerome: Dell has made a significant investment in software over the last few years and even now has a software division. While we have talked a little bit about appliances, most of our conversation has been about the software features that Dell now brings to the table. Can you walk me through some of the key components of Dell's software portfolio and the offerings it has?

Brett: Absolutely. One of the things I will tell you right off the bat is that Dell has *always* been in the software business. Before it might have been disguised in some of the different business units across PCs, servers, storage and networking, but what is great about the software division at Dell now is it shows that the company truly has a focus on providing end-

to-end solutions inclusive of software.

Creating a software division was a great way to consolidate the many software offerings the company had, and really focus on their development, marketing, sales, and most importantly, the integration of these products across the Dell portfolio.

It's ironic because when I first came to Dell 10 years ago, the server, storage and networking groups were a very healthy, very big part of the company, but a lot of customers did not know that at the time and only thought of Dell as a PC company. That's similar to what's happening with software today. Software is critical part of what Dell does and a very healthy part of the company. Dell has moved more and more in the direction of solutions. As it moves more so in that direction, toward becoming a more complete IT provider, software is playing and will continue to play a bigger and bigger role.

To me, software is really the glue that holds the solution piece at Dell together. We can provide management infrastructure for your server and PC environment. We can provide data protection and data recovery capabilities across your application and your storage environment. Software piece is a big part of that.

We break our software business into five key categories.

1. There's data center and cloud management.
2. There's mobile workforce management which is a big investment area with some new products coming out from Dell.
3. Our information management team is doing a lot of work around database, analytics and Big Data.
4. There is data protection, which is obviously a big focus.
5. Our security software.

All of the areas are what we would consider rapid growth areas

for us. They really provide a great solution story for our portfolio and/or the different products across Dell. We continue to get better at having our teams work with Dell's broader sales teams to provide the software expertise for customers who are looking for more software-centric solutions.

We have 6,000 team members globally, 1,600 software developers, and 2,500 sales teams that work with the broader sales organization at Dell. To give you an idea of the size and scope of Dell Software, 90 percent of the global 1,000 firms today are Dell software group customers.

This is something that people do not know very well ... yet. The reality is that Dell has been in the software business for a long time. It's certainly not a business we lack experience in. That said, I think right now we have a renewed focus on software and certainly a desire to grow this business and make it an even bigger part of Dell's end-to-end solution set.

In [Part I](#) of this interview series, Brett and I discussed the biggest backup and recovery challenges that organizations face today.

In [Part II](#) of this interview series, Brett and I discussed the imperative to move ahead with next gen backup and recovery tools.

In [Part III](#) of this interview series, Brett and I discussed four (4) best practices that companies should be implementing now to align the new capabilities in next gen backup and recovery tools with internal business processes.

In [Part IV](#) of this interview series, Brett and I discussed the main technologies in which customers are currently expressing the most interest.

In [Part V](#) of this interview series, Brett and I examine whether or not one backup software product can "do it all" from a backup and recovery perspective.

In [Part VII](#) of this interview series, Brett provides an in-depth explanation of Dell's data protection portfolio.

In [Part VIII](#) of this interview series, Brett and I discuss the trend of vendors bundling different but complementary data

protections products together in a single product suite.

TIBCO Event Processing: Relevant, Real-time Operational Intelligence

Deriving value from the plethora of unstructured data created by today's multiple sources of Big Data hinges on analyzing and acting on it in real-time. To do so, enterprises must employ a solution that analyzes Big Data streams as they flow in. Using TIBCO Software's Event Processing platform, enterprises can process Big Data streams while they are still in motion providing real-time operational intelligence so they may take the appropriate action while the action still has meaningful value.

Streams of Big Data Flowing In

Enterprises have more opportunities – and more reasons – than ever to capture multiple streams of Big Data coming in from numerous sources. Device sensors, Internet of Things (IoT), log files, RFID tags and social media platforms such as blogs, Facebook, Google+, LinkedIn and Twitter, all generate raw data that enterprises can utilize to make real-time assessments.

In this new Digital Economy the advantage goes to enterprises that can capture data, analyze it and then quickly and appropriately respond to it as events occur. Data's greatest value increasingly becomes the moment it is created or a short period of time (seconds, minutes or hours) thereafter. This makes it essential for enterprises to have a solution that can ingest and analyze this data, and in a timely manner, produce

the information that enterprises need to act appropriately to save money or turn a profit.

The Challenges of Extracting Big Data's Real-time Value

The ease and speed with which large volumes of data generated by machines and human activities are offset by the multiple challenges associated with quickly and effectively deriving value from this data. Specific challenges associated with extracting Big Data's value include:

- ***Ingesting data from numerous, different devices.*** Multiple bespoke protocols and industrial standards mean that little commonality exists in how device sensors, IoT devices, RFID tags and social media platforms transmit and receive data. This puts the onus on any data processing solution to account for how each of these devices or platforms transmits data so it may appropriately ingest the data and in the correct sequence.
- ***Storing and expeditiously processing the data.*** It is [estimated](#) that **50 billion devices** will be connected to the internet by 2020. Twitter already daily [averages](#) **500 million** tweets while Facebook [collects](#) approximately **500 terabytes** of data per day. Ingesting and analyzing the data from these sources in real-time time to derive value requires that any solution have the architecture and efficiency to keep up with these data rates.
- ***Establishing the data's context.*** Data arriving from each of these sources does not map into the traditional "name," "address," "email," and "phone number" fields used by relational databases. Rather data is created and stored in an unstructured format. This leaves it to the solution to establish the data's content and context as its meaning may change depending upon when and under what conditions the data was created

Fast Data Architecture Delivers Real-time Operational

Intelligence

Achieving operational intelligence requires a [Fast Data](#) architecture that analyzes Big Data in real-time as it happens. Big Data analytics were designed to look at historical information and produce analysis **after** the benefits associated with the collecting the data has passed. This “*Too Late*” approach makes it difficult if not impossible to reap the benefits of Big Data analysis for firms wishing to use and iterate on that analysis with live streaming data.

The TIBCO [Fast Data](#) architecture provides this missing link to realizing Big Data’s benefits. This architecture is designed around the processing, analysis and immediate insight into data in real-time. To accomplish this, it ingests and holds Big Data streams in memory as they arrive for a specified period of time. Holding this data in-memory expedites its analysis while also providing defined parameters to evaluate the data’s context.

Data held in memory is analyzed based upon one or more criterion to identify and spot patterns so decisions and actions may occur promptly and while there is still value, avoiding the Too Late architecture of existing systems and more recent Big Data models. For instance, a Fast Data architecture will:

- Continuously run queries against the incoming streams of data to determine if matching conditions exist to take action.
- Perform thousands or even millions of queries per second. Since new data is constantly arriving as old data ages, the conditions for whether or not to perform an action may change quickly.
- Provide the flexibility to add, change or remove queries as well as change the frequency of refresh rates as to how quickly queries are performed across the data in memory.

By continuously running queries against an ever changing set of data and then matching to real-time actions, solutions based on the TIBCO Fast Data architecture can finally deliver enterprises the operational intelligence that they need to take action while it still matters which saves money, improves customer satisfaction and drives profitability.

Fast Data's Real World Business Ramifications

Identifying and creating new revenue opportunities, improving operational efficiencies and driving down CAPEX and/or OPEX costs are just some of the possibilities that result from implementing a solution based on the TIBCO Fast Data architecture. By analyzing large amounts of data created within a defined period of time, then creating and executing queries based on business rules against that data in real time, enterprises can drive customer satisfaction and revenue in new and innovative ways.

For example, more and more people have Internet-connected mobile devices. .By associating the device with the customer and their location the TIBCO Fast Data enabled solution can determine if the individual is a new or returning customer and potentially even pull up past purchases made by that individual. Using that information, an email or text may be sent to that individual's device that contains a coupon, offers a deal that is only valid while they are in the store or recommends a item to buy based upon a prior purchase they have made. This is context-aware marketing and customer service that adds value.

Enterprises may similarly leverage the TIBCO Fast Data-solution to improve operational efficiencies and/or drive down costs. For example, delivery services can improve efficiency with effective rerouting, and minimize staffing by predicting and updating package volumes in real time. Real-time information can also enhance partner relationships and uncover new business opportunities. And the same fast data platform

can simultaneously improve customer experience by exposing more information about real time locations of package and predictions for when they will be delivered.

TIBCO Event Processing: Relevant, Real-time Operational Intelligence

Enterprises have had access to multiple streams of Big Data generated by external social media platforms, mobile devices and IoT as well as internal sources such as device sensors and RFID tags for some time. Yet maximizing value from historical analysis of past data and live streaming data typically requires analyzing and acting upon it in minutes or even seconds after its creation.

TIBCO [Event Processing](#) provides enterprises the TIBCO Fast Data architecture that they need to do real-time processing and analytics. By quickly ingesting and analyzing data and then making real-time decisions based upon it, TIBCO Event Processing gives enterprises access to the operational intelligence that they need to make an informed business decision based upon the best data available in the environment in which they operate.

Preview of the 2013 DCIG Private Cloud Storage Array Buyer's Guide

Around two years ago the DCIG 2011 Enterprise Scale-Out Storage Buyer's Guide was released. At the time we mentioned that scale-out systems were being used to store "Big Data" and create private storage clouds. Since then scale-out storage

systems have become the foundation for building out private storage clouds which prompted DCIG to change the name of our refreshed Buyer's Guide to better reflect the intended use case for these storage arrays.



While the term "*private storage cloud*" is certainly hip and may be used a bit too liberally in today's culture, this label is actually much more descriptive of what scale-out means for an end user rather than what it does technically for storage engineers.

In the last few years "the cloud" has moved from being a quasi-technical term to going main stream. A CFO's executive assistant probably does not know what a "*scale-out storage system*" is or does. However she has probably heard of the term "*storage cloud*" and grasps at a high level what it can do for her company. By renaming this Buyer's Guide to "Private Cloud Storage Array," DCIG helps to better communicate how scale-out storage arrays are best deployed and leveraged in businesses.

That said, private cloud storage isn't about ubiquitous access to users documents and pictures on every imaginable device. The savvy reader who will soon have the new Buyer's Guide in their hands understands that the arrays presented in the DCIG 2013 Private Cloud Storage Array Buyer's Guide are the building blocks for many different storage needs, most of which will never directly touch end users.

DCIG recognizes that many organizations will continue to view private clouds as a natural extension to their archiving and backup solutions. However, we believe that as this category matures organizations are becoming more willing to place mission critical data onto their cloud storage. To that end

the refreshed Guide places more focus on management and other features that reflect the diverse ways scale-out storage arrays are being utilized. In particular, we've added a new category on VMware vSphere integration. We also have placed additional emphasis on features such as SSDs and other Flash technologies both for raw storage and use as cache.

One of the reasons we believe that this space is maturing is the emergence of products that are focused on solving specific problems. DCIG was very encouraged to find an all flash array from [Nimbus Data](#) Systems – its [E-Class](#) Flash Memory array – appearing in this Guide.

At the other end of the spectrum is [Gridstore](#). While owning the lowest price point in the survey it has provided a very unique value proposition for small enterprises that need simple, reliable, and resilient storage. By using a software driver that is installed on both servers and clients Gridstore moves the traditional controller out to the data consumers and off the storage device itself. While this limits the product's use of more "traditional" storage it provides a useful option for many small enterprises.

Another unique entrant was [Coraid](#). Coraid uses ATA-over-Ethernet (AoE) to provide a solution that should provide advantages in environments that require multi-pathing and high resiliency. Other unique products include object storage options from [Cleversafe](#) and [DirectData Networks](#), a virtualization-in-a-box solution from [Scale Computing](#), and a video and media focused scale-out [solution](#) from [Pivot3](#). We also include more archiving focused products from [IceWEB](#) and [Overland Storage](#).

This is not to say all the arrays we researched were focused on specific problem sets. Well known vendors such as [Dell](#), [EMC](#), [HP](#), [IBM](#) and [NetApp](#) still provide several models intended for more general use. These models generally include more advanced features not present on products focused on a particular role.

DCIG was pleased to find the new Guide bore out at least one area we said we expected to see growth in the DCIG 2011 Enterprise Scale-Out Storage Buyer's Guide. That Guide included only two vendors who supported deduplication, NetApp and Symantec. This Buyer's Guide now includes seven vendors and ten models that support deduplication while Symantec no

longer has a qualifying product. We further note that many of the vendors who do not offer deduplication are not positioning their private cloud storage arrays for the purpose of archiving. DCIG is encouraged by this trend.

Some other interesting statistics about the arrays in the Guide include:

- 62% support at least 1 PB of raw storage capacity for the entire cluster
- 60% support some portion of VMware's VAAI (vStorage API Array Integration)
- 59% of the models supported at least 48 TB of raw storage per node
- 48% support some form of flash technology for raw storage
- 38% of models that support NFS include support for NFS v4.1
- 32% support 8 Gb FibreChannel
- 1 model claims a tested configuration of over one (1) exabyte. Several other models have theoretical limits of over 1 XB

Stay tuned as the release of the DCIG 2013 Private Cloud Storage Array Buyer's Guide is just around the corner.

Most Read Blog Entries on DCIG's Web Site in 2012 – Honorable Mention

Mention the year 2008 or 2009 to almost any person and it almost inevitably elicits a negative reaction in terms of how those years were from a business perspective. However as DCIG renews its annual tradition of reflecting back on what blog

entries were most read on its website during the course of 2012, 2008 and 2009 emerged as very good years in terms of DCIG providing content that is still relevant and frequently read in 2012. Today and over the next four (4) business days, I will share what blog entries garnered the most attention on DCIG's website in 2012.

I use a very simple metric to evaluate which blog entries garnered the most attention each year: I simply look at the total number of page views on a specific blog entry over the course of a year. What is noteworthy about this year's results is that even though most of the blog entries that topped this year's results were written in 2008 and 2009, they covered topics that organizations of all sizes are now addressing and/or researching in 2012.

So to kick off this year's annual recap, I begin with those blog entries that garnered a lot of attention but did not quite make it into DCIG's Top 10 most read blog entries for 2012. However to ignore them would be doing these blog entries a disservice so I am awarding the blog entries listed below the status of "**Honorable Mention**" for doing so well but coming up just a bit short in DCIG's year end results.

Honorable Mention

(in alphabetical order)

- **A "Destroy All Data" Policy Will not Absolve You of Your Data Retention Responsibilities** ([link](#)). Howard Haile. This blog entry written in June 2009 is one of many written in 2008 and 2009 on the topics of Big Data and eDiscovery that struck a chord with DCIG readers in 2012. Companies already face increasingly complex and difficult decisions when it comes to managing their Big Data stores and even hate to think about the implications of eDiscovery. While some are adopting a "destroy all data" policy, this blog entry covers why this is **NOT** a recommended course of action.

- **Cloud Storage Architecture Gives MSP New Found Flexibility to Respond to Spikes in Application Requirements** ([link](#)). *Jerome Wendt*. These days everyone loves to talk about cloud storage and the benefits it provides but in 2009, cloud storage was still on the drawing boards for most companies. Not so with the MSP covered in this particular blog entry. It brought in HP 3PAR's private cloud storage architecture so it could do more than easily scale to meet increased demands for capacity. It also gave it the flexibility to meet an annual spike in performance associated with a client's charity event.
- **Deciphering Application Metadata is Data Deduplication's Next Frontier** ([link](#)). *Jerome Wendt and James Koopman*. Deduplication is now an integral part of the backup process for many companies. However based upon the increased interest in this 2009 blog entry in 2012, it suggests that companies are looking to optimize how they deduplicate their backup data. In this blog entry, James and I discuss the need for backup software and deduplication appliances to look into the backup stream itself to better deduplicate data.
- **Four Data Megatrends for the Decade of the Teens** ([link](#)). *Jerome Wendt*. This blog entry got a lot of attention in 2011 when it was published as it was picked and re-published on a couple of DCIG's partner sites. However the interest in the four (4) datacenter megatrends covered in this blog entry continued into 2012 as companies continue to centralize their infrastructure administration, automate the datacenter, look to satisfy increasing end user needs for instant gratification and put a cloud infrastructure in place.
- **Using Virtualization to Anticipate and Solve Today's and Tomorrow's VMware Virtual Machine Backup Problems** ([link](#)). *Jerome Wendt*. Server and storage virtualization were already hot in 2009 but the challenges of doing backup in these newly virtualized environments were just

beginning to be understood. It is clear that based upon the large number of people reading this blog entry in 2012 that more companies are experiencing the pain associated with doing backups in their virtualized environment and are looking for ways to do more than just solve their backup pain but identify the source of it before it impacts them.

Check back on DCIG's web site tomorrow when I cover the most read blog entries on DCIG's web site that were written in 2012.

Tips to Choosing the Right Tape Library for Use with Big Data and Cloud Initiatives (Webcast)

The factors that influenced which tape library to use in your environment used to be much simpler in nature when tape was used primarily as a backup target. But as disk has evolved to assume that role, tape libraries have evolved to provide new features so they may assume a much more strategic position within organizations to support their Big Data and Cloud initiatives. In this webcast, I take a look at how to choose the right tape library for your environment in light of these new forces that are impacting the use of tape libraries within organizations.

In this webcast, I discuss the following:

- ***Challenges that every organization faces when making a tape library buying decision.*** Buying a tape library today is much different proposition than it was five or ten years. There are over 60 products from which to chose, there are substantial differences and similarities between these products and many have doubts about tape's viability going forward.
- ***Why tape's operational costs are becoming a larger factor in the choice of this media for hosting Big Data and data stored in the cloud.***
- ***The role of tape libraries in meeting the specific requirements of Big Data and the Cloud.***
- ***How these factors influenced which tape libraries were included in the recent DCIG 2012 Big Data Tape Library Buyer's Guide.***
- ***How tape libraries were scored and ranked in the Buyer's Guide***
- ***Recommendations on how to best use the DCIG 2012 Big Data Tape Library Buyer's Guide***

You may register to download an entire copy of the DCIG 2012 Big Data Tape Library Buyer's Guide at this [link](#).

Virtual Federated Data Stores are an Attractive Alternative to Massive Central Data Stores

An integrated and centralized data store model that enables stakeholders from throughout an organization to harvest and

analyze data on the same platform continues to be the goal of many organizations from the Global 2000 as they strive to address the requirements of Big Data. However, with today's decentralized and cloud based storage systems and data storage requirements for the Global 2000 reaching the petabyte and even exabyte levels, massive centralized single data store infrastructures with single points of failure such as Apache Hadoop may not be the most effective long term solution.

Driven by social media, mobile computing devices and cloud computing, the volume of Electronically Stored Information (ESI) is increasing at an accelerating rate. As a result, organizations are not only faced with the unprecedented challenges in managing the costs and legal and compliance requirements associated with this increase in data, they are also now under pressure to analyze all of this new data to uncover competitive advantage for product marketing, sales, Enterprise Resource Management (ERM) and even Human Resource (HR) management.

The rewards for successfully managing and leveraging what is now known as Big Data are well documented. At IBM's Smarter Analytics event in March 2012, clients and partners presented success stories about how organizations are driving business value out of big data, analytics, and IBM Watson technology.

Examples included:

- City of Dublin, Ireland, using thousands of data points from local transportation and traffic signals to optimize public transit and deliver information to riders.
- [Seton Healthcare](#) mining through vast amounts of unstructured data captured in notes and dictation to get a more complete view of patients. Seton currently uses this information to construct programs that target treatments to the right patients with a goal of minimizing hospitalizations in the way that most

efficiently optimizes costs with benefits. The ability to mine unstructured data gives a much more complete view of patients, including factors such as their support system, their ability to have transportation to and from appointments, and whether or not they have a primary care physician.

- [WellPoint](#) using IBM's Watson technology to improve real-time decision-making by mining through millions of pages of medical information while doctors and nurses are face-to-face with patients.

Unfortunately, as organizations realize they have to do "*something*" to survive Big Data and therefore begin the journey through the Big Data Maturity Model, many inadvertently field redundant efforts at the department, division and other sub-corporate level.

As an example, it wouldn't be unusual for an organization attempting to address Big Data issues to have the following groups performing the same basic data harvesting, remote storage and analysis functions with different yet very similar technologies:

- A corporate Business Intelligence (BI) group within the Information Technology (IT) division to harvest and analyze data for internal business consumers
- A corporate Governance, Risk and Compliance (GRC) group associated within the office of the Chief Financial Officer (CFO) to harvest and analyze data to reduce risk and ensure compliance
- An eDiscovery group within the legal department to harvest data for computer forensics and Early Case Assessment (ECA)

What organizations should be striving for as they evolve up the Big Data Maturity Model is an integrated and centralized data store model that enables stakeholders from BI, GRC, eDiscovery and other groups as required to harvest and analyze

data on the same platform.

Apache Hadoop, an open-source Big Data platform for developing and deploying distributed, data-intensive applications has become a attractive solution among the Global 2000 for accomplishing this goal. However, with today's decentralized and cloud based storage systems and data storage requirements for the Global 2000 reaching the petabyte and even exabyte levels, **massive centralized single data store infrastructures with single points of failure such as Hadoop may not be the most effective long term solution.**

The alternative solution is to match today's decentralized; cloud based virtual environments with **virtual federated data stores**. Technology vendors such as VMware are already investigating solutions.

A recent article posted on CRN.com [states](#), "*As server and storage virtualization become standard in the data center, Apache Hadoop and software defined networking are looming as VMware's next big challenges.*" Earlier this year, VMware launched an open-source project called [Serengeti](#) that includes a free deployment toolkit for deploying a Hadoop cluster on vSphere.

Other Big Data software startups are also addressing the requirement to provide a virtual data store with what I would characterize as Big Data middleware. One such vendor is [Tarmin](#) with its [GridBank](#) virtual software solution. GridBank enables users to create a virtual object based federated data store and provides a single management console to monitor and analyze rapidly growing, geographically dispersed unstructured data repositories.

An integrated and centralized data store model that enables stakeholders from BI, GRC and eDiscovery to harvest and analyze data on the same platform continues to be the goal of many organizations from the Global 2000 as they strive to

address the requirements of Big Data. However, with the rapid increase in the amount of data and today's decentralized and cloud based storage systems, Virtual Federated Data Stores may be an attractive alternative to massive central data stores.

Five Initial Steps to Take to Meet the Governance, Risk and Compliance Obligations Brought on by Today's Big Data File Stores

The accelerating increase in the amount of unstructured Electronically Stored Information (ESI) is leaving IT organizations struggling with how to store and manage all of this new information. Aside from needing to provide the underlying storage infrastructure to host this amount of data, companies are also faced with the task of properly managing their Big Data file stores to meet both existing and emerging governance, risk and compliance (GRC) obligations. To do so, there are five initial steps they can take now to get their organization in front of these demands.

According to a 2010 [report](#) by IDC, the amount of information created, captured or replicated has exceeded available storage for the first time since 2007. The size of the digital universe this year will be tenfold what it was just five years earlier. According to this same IDC report, the volume of unstructured ESI is expected to grow at over 60% CAGR (Compounded Annual Growth Rate).

Forrester Research adds some teeth to this. As [reported](#) in an article that appeared on Forbes website last week:

- The average organization will grow their data by 50 percent in the coming year
- Overall corporate data will grow by a staggering 94 percent
- Database systems will grow by 97 percent
- Server backups for disaster recovery and continuity will expand by 89 percent

Overseeing the expansion of storage space and ensuring that the data is protected has become a minor part of the overall task of Big Data file storage and management. Business stakeholders and the Information Technology (IT) organizations from enterprises of all sizes and across all industries must now comply with a growing list of GRC regulations or face potentially fatal financial penalties to the enterprise.

The most obvious laws to which they are subject include:

- Sarbanes-Oxley ([SOX](#))
- Health Insurance Portability and Accountability Act ([HIPAA](#))
- Gramm-Leach-Bliley ([GLBA](#))
- Federal Information Security Management Act ([FISMA](#))
- Consumer Information Protection Laws
- Federal Rules of Civil Procedure ([FRCP](#))

Further, the list of new regulations is growing. The passage of The Patient Protection and Affordable Care Act ([PPACA](#)) will result in the US Government **adding 159 new agencies, programs, and bureaucracies** to assist with the compliance of **over 12,000 pages of new regulations**. Over the past ten years, in response to the threat of international terrorism, the US Department of Homeland Security ([DHS](#)) has added hundreds of new regulations. Finally, cyber terrorism, including acts of deliberate, large-scale disruption of enterprise computer networks, is now a

reality that all businesses must face.

In the face of this, Big Data file storage and management vendors, along with the associated industry consultants, have developed a list of hardware and software requirements and associated value propositions to help enterprise buyers decide which Big Data file storage and management platforms to purchase.

Yet before they buy, there are five steps that buyers should first take to ensure they are prepared to meet the GRC obligations brought on by today's Big Data file stores:

- **Internal Collaboration:** File management and GRC requirements affect business stakeholders from the boardroom to IT to the manufacturing floor and loading dock to the accounting office. The development of cross functional workgroups and the promotion of internal collaboration between functional experts is the key to successfully identifying, understanding and addressing all of the requirements and issues involved in Big Data file management across the entire enterprise.
- **Network Architecture Planning:** Over the past 25 years, enterprise architectures grew with little or no planning resulting in wasteful redundancy and little or no access to all the enterprise data as may be required to comply by today's GRC requirements. The advent of the Internet and now cloud computing has exposed these decades of poorly planned networks resulting in them become more of enterprise liability than an asset. The time is now for IT to hit the restart button and explore new options such as virtualization, hybrid cloud architectures and the use of cloud service providers (CSPs) that enable them to better leverage, manage and optimize their existing infrastructure.
- **Security:** The introduction and proliferation of

portable storage devices, Wireless Internet, mobile computing devices, enterprise Software-as-a-Service (SaaS) applications, cloud storage, blogs and social media such as Facebook, LinkedIn and Twitter, data theft and cyber attacks are a real issue for which many (and arguably most) companies do not have a good answer. Now is the time for IT to take a serious look at their internal file access policies and move quickly to address any existing shortcomings.

- ***Data Retention Policy Development and Implementation:*** Sarbanes-Oxley (SOX), the Health Insurance Portability and Accountability Act (HIPAA) and the Federal Rules of Civil Procedure (FRCP) all have very specific data retention guidelines for what types of ESI data an enterprise has to keep and how long to keep it. Enterprises must investigate and document these requirements, development data retention policies and acquire the appropriate software to ensure compliance.
- ***Technology Vendors and Consulting Partners:*** Business stakeholders and IT management may be overwhelmed with the task of addressing the issues of successfully meeting the GRC obligations of big file storage and management. If this is the case, reach out to the hardware and software vendor community and ask how their solutions support these issues. If required, engage the services of vendor independent consulting partners to act as trusted advisors to assist in the successful navigation of the required cultural transitions and the acquisition of the best technology platforms.

The accelerating increase in the amount of unstructured Electronically Stored Information (ESI) has put IT on the defensive as it grapples with how to store and manage all of this new information. The traditional approach of simply “buying more,” overseeing the expansion of available capacity and then ensuring that the appropriate backups are completed

are woefully inadequate as these tasks are about to take a back seat to the much larger issues that Big Data file management creates.

Business stakeholders and IT need to act now to bring their infrastructure under control so they can get in front of the growing list of existing and emerging GRC regulations to which they are subject. By following these five steps outlined above, enterprises will put themselves in a position so that when they purchase a product, they will have a good grasp of what their true enterprise challenges are and increase their likelihood of bringing in a product that addresses them.

A Mid-Year Look at the Top Technology Trends So Far in 2012

To say with any degree of certainty what technologies will be hot in the next 6 – 12 months generally takes equal amounts of smarts and industry insight coupled with a little bit of luck sprinkled in to get it right. So as I compare what I forecast earlier this year to what I see taking place now, I was certainly right on some points but premature in predicting others. So today with the midpoint of 2012 upon us, I thought I would take a look at the five specific technology trends impacting organizations right now.

Two forecasts from a [blog entry](#) that I wrote back in January 2012 are shaping up to be for the most part on the money.

The first prediction had to do around the **growing use of backup appliances in 2012**. If anything, end user desire to

acquire, deploy and use backup appliances has exceeded even my admittedly optimistic viewpoint earlier this year about their adoption.

As I stated at the time, the consumerization of IT is well under way in all size organizations and shipping backup software as an appliance does a great deal to simplify its deployment and implementation.

Contributing to this trend, organizations are consolidating IT staff onto single teams (75% of organizations already have a single team involved in all IT decisions according to the [results](#) of one survey.) Backup appliances make it possible for organizations to accomplish this and still effectively deliver on the tasks for which they are held responsible.

The other trend I picked up on in early 2012 was the **shift in focus on the part of organizations from backup to recovery**. Backup appliances have certainly contributed to this trend as IT administrators can quickly setup and configure backup software in their environment so that the next time they interact with it may well be in the context of recovery.

However an equally large contributor to this focus on recovery is the growing use of array-based snapshots. As was shared at the recent [Dell Storage Forum](#), a recent analyst survey found that 52% of organizations now use array-based snapshots as their primary way to protect production applications. This has led to easier, faster recoveries driven in large part by software products like CommVault [Simpana](#) and Symantec [NetBackup](#) that have supporting snapshots a specific focus.

In terms of new trends did I not comment on in January 2012 but that I now see emerging in mid-2012, there are three of them.

First, Big Data, the Cloud and Virtualization are the new corporate realities. These three concepts are no longer

emerging trends but now influence in some way how almost every organization makes every technology buying decision. While there are certainly nuances in terms of how applicable each one is to each size organization (*for instance, large enterprises may deal with "Big Data" while small businesses grapple with "Bigger Data",*) the influence of these three concepts is real and undeniable.

Second, servers and storage are becoming more tightly integrated. While a number of storage arrays already offer flash as either cache or solid state drives (SSDs), more storage vendors are looking to move flash closer to or into the server. Once there, they are then looking to manage the server flash in a manner akin to how they manage data placement and storage capacity on their arrays.

Third, virtual infrastructure management is begging for a software solution. As much as everyone loves virtualizing the infrastructure, in the same breath they express how difficult it is to manage and troubleshoot. This is why a software solution that is specifically tailored to monitoring, managing and even fixing trouble spots such as do root cause analysis within virtualized environments needs to emerge sooner rather than later.

Clearly other technology trends beyond the five mentioned above are gathering momentum in mid-2012. However I highlight these because all of them are largely a direct result of IT staff in organizations being pushed to do more productive using the same or fewer budget and resources. So while technologies like server-storage integration and virtual datacenter infrastructure management are still in their infancy and maturing, this is where vendors are currently putting many of their efforts and where organizations should expect to see the most innovation in the coming 6 -18 months.

Boosting Transactional Performance with the Fusion-io SDK Kit; Interview with Fusion-io Sr Director Product Mgt Brent Compton Part II

The deployment of flash memory storage as either storage or memory almost inevitably results in increases in application performance. However to get the real 'kick' in performance that today's transactional applications need and which flash can provide, a more elegant approach to flash's deployment is needed. Today I continue my discussion with Fusion-io Senior Director of Product Management, Brent Compton, who elaborates on the APIs that the Fusion [ioMemory SDK](#) exposes that make this boost in transactional performance possible.

Ben: *Direct access to flash RAM is definitely a foundational requirement. The ability to offload processing of key-value stores is another basic yet consistently needed functionality for developers. What else can we expect?*

Brent: First, I'd like to distinguish between a 'primitive' and an 'API' in our SDK nomenclature. A primitive is a single, foundational interface call while an API is a family of related interface calls. For instance, the directKey-Value Store API is a family of related interface calls.

One of the primitives is the **Atomic Multi-block Writes**. It takes advantage of one of the native properties of our ioMemory flash translation layer: the log structured write

mechanism. This mechanism provides a basic copy-on-write foundation.

Just a simple illustration, if you write three blocks A, B and C and then you come along and update A, unlike a disk which performs a rewrite in place, ioMemory writes 'A' to the tail of the log. So both A and newly-written 'A' exist on ioMemory.

There are many different ways that applications could exploit this foundational copy-on-write property, not the least of which is to provide atomicity for multi-block writes. This occurs when an application says, *"I need to write a bunch of blocks and I need to ensure that all of them are written, or none of them are written."*

A practical example of this might be writing a combination of data and metadata. Both the data and corresponding metadata need to be written to ensure integrity of the update. If only part of the data is written, or only the metadata is written, the data repository will be out-of-sync. All written, or none written.

This means we have all the makings of a double buffered write. If there is any interruption to a multi-block write we have all the mechanisms in place to roll back to the previous content. We just ignore a partial write of those blocks as if it had never occurred.

We gave a sneak preview of **Atomic Writes** in October 2011 at [Oracle Open World](#). In conjunction we pulled down some of the MySQL open source for the InnoDB storage engine.

We modified MySQL InnoDB by replacing its double buffered writes with native calls to our atomic multi-block write interface. We saw significant latency reductions, performance improvements with more IOPS, while reducing the code path. So it's one of those *"less code, better performance"* stories.

Ben: *I think those are some good examples. What other*

categories of APIs are available?

Brent: The next category is the **memory access API family**. Note the different use of words: memory access API versus direct IO API.

Flash has always been a hybrid of memory and storage but to date, its uses in industry have been storage. We're offering the industry's first memory access semantics to flash. There a couple of different APIs under this memory access API family. I'm going to highlight one.

On January 5th [Fusion-io demonstrated one billion IOPS](#) using a technology we call Auto Commit Memory (ACM). This SDK becomes the vehicle through which we provide this capability to software developers through an **ACM API**.

The essence of it that is an application can designate a region of its process virtual address space as persistent. ACM provides API semantics to attach this persistent region to Auto Commit Memory, to ioMemory, such that anything written to that region of the process virtual address space is guaranteed to be automatically persisted.

This is a mechanism has some similarities to mmap, so you have the benefit of saying, *"Wow! I can eliminate a lot of my code complexity by persisting my data, without having to resort to IO primitives. I can just store data in that location of memory and have somebody else worry about persisting."* In this case the 'somebody else' is the ioMemory SDK.

However, unlike mmap, and this is key, **ACM guarantees persistence**. In other words when you write something to memory allocated with mmap, and there is an interruption of service, you're not guaranteed that it will be persisted. You don't have the durability of writes which means you have to resort to various other mechanisms of journaling or double buffered writes or things like that.

On the other hand, when you write something to **Auto Commit Memory**, by design it will be automatically committed. In other words, ***it is durable across service interruptions such as power failures.***

Note that part of the ACM API will be a write barrier operation, like a flush, ensuring that the data is cleared from the processor complex, various levels of CPU caches and what not. Once flushed from the processor complex, it's automatically persisted to ioMemory.

What attracts a lot of database developers to this new API is the notion of solving the tail-of-their-transaction log performance inhibitor. By definition, it is the transaction log through which they can ensure ACID properties of transactions.

Previously developers had to issue blocking synchronous I/Os at the tail of their log, to ensure that the most recent writes before service interruption were durable. With our ACM API they can convert that blocking synchronous IO to a non-blocking asynchronous IO by maintaining the tail of their transaction log in auto commit memory.

They may still persist the tail of their log to a backing store but they will not need to do it synchronously through a blocking IO. If there's an interruption, for instance upon a system or an application restart, they can always recover their state through what was persisted in auto commit memory. So developers are quite keen on that.

In [Part I](#) of our interview series, we discussed how the Fusion-io SDK kit will help to unleash the next gen properties of flash.

In [Part 3](#) in this interview series we will discuss the "DirectFS" API, a native POSIX compliant direct file system layer and discuss the more technical aspects of how the SDK works.

In [Part 4](#) in this series Brent and I discuss the semantics of using the API in the C language and how Fusion-io is leveraging its early access partnerships.

Fusion-io SDK Kit Unleashes Next Gen Properties of Flash; Interview with Fusion-io's Brent Compton Part I

Last week developers of enterprise applications got some new toys to play with in the storage memory realm. The newly released ioMemory SDK will grant developers the ability to better utilize the potential of Fusion-io's line of enterprise flash memory storage. Fusion-io expects the SDK will simplify code bases while providing a sizable performance boost.

We begin our discussion with Brent Compton, Senior Director of Product Management with Fusion-io. The first entry in the series will focus on the need for the new SDK and begin to discuss some the functionality developers should expect.

Ben: *Brent, I'm looking forward to this interview because I get to wear both my administrator hat and my developer hat! To start off with, can you please tell me a little about what you folks have been doing up to this latest announcement?*

Brent: In short, [Fusion-io](#) has pioneered a next generation flash memory storage platform for enterprise data centers. Our flagship product, [ioMemory](#), is a hardware-software combination which provides enterprise grade high performance flash memory for enterprise apps.

Last week we [announced](#) our ioMemory software development kit ([SDK](#)), which will enable applications to run natively on flash, or natively on ioMemory, for the first time.

You may have read academic white papers over the last couple of years talking about **providing native access to this flash memory tier**. As far as we're aware, **we'll be the first ones to offer that through this software development kit**. Before we get into what will be provided to developers, let me give some brief history for context.

As you are aware, around 1956 IBM defined a half century of I/O with a single word, seek, when they invented the disk drive. To exploit this new hardware innovation, they also gave developers a primitive to access data randomly.

Random access to data transformed software development, which before had been working with sequential access devices. In the half century that followed, virtually all operating systems, databases, and applications have been playing to that tune. In other words, they have been built to work with, and more recently, sometimes to work around, that rotational latency.

When flash memory first came on to the scene as an enterprise grade device in 2007 or so, it was disguised to look like a disk.

This was an important first step. As enterprises learned to trust the medium and learned where to use it, it was important that it be offered transparently and easy to integrate. So **the industry disguised flash memory to look like a disk for ease of adoption**.

Now that's where Fusion-io's path diverges. Unlike the other vendors who are classically described as "SSD vendors," we've always described our media as ioMemory. From the beginning we chose a fundamentally different architectural path, looking forward to the day when this would be a new tier, a hybrid of memory and storage.

And yes, while the important first stage of adoption was to disguise it to look like a disk, it's now mainstream and obviously a great deal of market enthusiasm surrounds the adoption of flash memory and flash solutions in general. With flash memory deployments now commonplace, ***it's time for us to crack open and exploit some of the native underlying properties of flash which have been effectively hidden to date.***

Returning to the comparison with the disk drive where we began, the parallel we like to use with the new ioMemory SDK is that when application I/O converted from sequential access tape to random access disk, a couple of new programming primitives were exposed to allow developers to exploit that hardware innovation. Likewise, to tap into the native properties of flash memory hardware/software innovation, a new set of primitives, or APIs, need to be exposed to developers for them to harness and exploit those native characteristics.

Ben: *Let's start talking about that. What functionality should developers expect from the new SDK?*

Brent: There are four key capabilities that developers will be provided. ***The first capability will be direct-access I/O.*** Direct-access I/O enables developers to bypass file systems the kernel block layer, and other I/O layers tuned over decades for disk drives. Application I/O is plumbed directly through to the ioMemory device.

The ***directKey-Value Store API*** is an example of APIs found within the direct-access I/O family. One of the native characteristics of the ioMemory flash translation layer is that it is natively a key value store. Everything that it stores inside of its log append structure are a key value pair. ***It stores a logical block address together with the associated data with every single write.***

As you're aware, the world is abuzz with activity right now in

the whole unstructured data, NoSQL, key value store solution space. All kinds of vendors new and old are building solutions for that market. We are already working with a number of them.

Minimally, we enable them to shorten the code path when performing key-value store I/O operations by eliminating duplicate logic in their code and in the ioMemory software layer. Maximally, we actually reduce their code because we're doing some of the low-level heavy lifting for key value store implemented natively on flash, exporting to them a key value store API interface.

Using this directKey-Value Store API, they can effectively pass in the key and the value, and we take over from there providing the hashing, collision management, and native storage and retrieval of key-value pairs. So it **offloads a lot of complexity from their code so they can focus on their value add** without doing some of the native stuff which we can do better, faster, cheaper, being closest to the hardware.

In [Part 2](#) of this interview series Brent will continue to discuss the primitives that developers will have access to including atomic multi-block writes. We also discuss how familiar the API will be to developers.

In [Part 3](#) in this interview series we will discuss the "DirectFS" API, a native POSIX compliant direct file system layer and discuss the more technical aspects of how the SDK works.

In [Part 4](#) in this series Brent and I discuss the semantics of using the API in the C language and how Fusion-io is leveraging its early access partnerships.

DCIG Interactive Buyer's Guide (IBG) Delivers First of a Kind Research as a Service for Enterprise Technology Buyers

We live in the information age where data is being produced at rates that almost boggle the mind. But living in the age of Big Data does not translate into this data being easily available and digestible. This is especially applicable when it comes to accessing, sorting and generating reports on information about products that organizations need to make informed buying decisions.



The new DCIG Interactive Buyer's Guide (IBG) fundamentally addresses this basic organizational need. Delivering information about enterprise technologies in the form of Research as a Service, the IBG quickly and easily enables organizations to:

- Access the information that they need about the products and features being considered
- Select the features that they need in their environment
- Produce needed reports to justify technology buying decision

Over the last few months DCIG has been [documenting](#) both the inadequacies of current product information (vendor data sheets, analyst reports, articles, etc) and the struggles that organizations face in digesting this information. This information provides:

- Detailed information about a specific product or product features; or
- Technology overviews that explain where a specific product fits in the market

Further, these static reports tend to get out of date and require users to filter through the information contained in them to identify which products possess the features that are the most important to them. As a result, users are forced to spend a great deal of time understanding what products are available, what features they possess and then make a technology decision based upon their research. Further, I have seen many (including myself) create very complicated spreadsheets to track which features are supported on each product.

The DCIG IBG changes all of that. Now organizations have at their fingertips both the raw product information they need ***AND*** the tools they need to easily view, filter and print this information. Maybe most importantly, they can access this information ***WITHOUT*** needing to create a highly complex spreadsheet in order to arrive at their conclusions.

Here are the key benefits that the IBG provides:

- ***Access to objective product information.*** Support for specific product information is based directly on information found on the vendor's web site, product literature, and a completed survey. DCIG takes great effort to make sure the most accurate product information is presented in these guides absent any

bias.

- ***Up to date information.*** As new features are added to existing products or new products are released, the IBG can be quickly and easily updated to include this new product information so the product features can be accurately assessed against other products in the market.
- ***Side by side comparisons.*** The DCIG IBG enables organizations to do side by side of comparisons of up to five products and as many features (up to 100 or more) or as few (1) as they wish to compare. In this way organizations can quickly compare the features that really matter to them based upon their requirements.
- ***Branded reports.*** All reports are branded with a DCIG IBG logo and date and time stamped in an appealing, professional format that can be used internally or shared with trusted partners. These reports become invaluable for those individual who need to present highly technical information in an appealing format to executives or those responsible for making buying decisions.
- ***Social networking.*** The DCIG IBG includes a section for comments on each product so as organizations test or evaluate different products, they may leave their comments in this repository rather than risking that they are lost or get buried Word docs, emails, or spreadsheets. Further, only individuals within the company with access to the IBG can access those comments preserving the intellectual property of each company.

The DCIG IBG achieves an important milestone in delivering a new type of analyst report. Rather than organizations having to rely upon information that is dated, static or hard to manipulate, they now for the first have access to information in real time and can take advantage of the best of what

analysts, journalists, vendors and even their own company have to offer and puts it into a portal that they can use.

The DCIG IBG is available immediately with pricing starting at \$5,000 per seat. Modules currently available include the DCIG 2012 Midrange Array Buyer's Guide and the DCIG 2011 Small Enterprise Storage Array Buyer's Guide. Please reach out to DCIG's VP of Business Development, Jim Nash, at jim.nash@www.d cig.com or you may call him at [1.844.324.4552](tel:1.844.324.4552) (844.DCIGLLC).

DCIG Launches First Interactive Buyer's Guide for Buyers and Providers of IT Storage Products – ([/ibg](#))

DCIG invites end users to participate in a freemium version of the Interactive Buyer's Guide (IBG). Please go to <https://ibg.d cig.com/freemium> to signup for FREE access to the Interactive Buyer's Guide for 30 days.

The press release associated with the IBG may be found [here](#).

The Days of Backup Software Being 'One Size Fits All' Are Over; Interview with Quest Software Sr VP Walter Angerer Part II

Today's expectations for always-on environments, coupled with the introduction of Big Data into enterprise environments are stretching the capabilities of today's backup software well

beyond what it was ever intended to solve. As such, enterprises can no longer look at backup software as a 'one size fits all' approach.

However, the change in perspective goes beyond even that. Vendors and users alike must change their view of backup and recovery from an infrastructure-centric view to a server-centric view. Today, I continue my interview with [Quest Software](#) Senior VP Walter Angerer, as we discuss the new directions that backup software must take in the enterprise.

Jerome: *One of the things I see happening is that we are definitely on the road to recovery. The long standing problem of creating a good, initial copy of the data has largely been solved. As a result, I am seeing a turn in my conversations with end-users and vendors alike to topics such as recovery, search, or testing and development using that good copy of data. How does this change how backup data is managed and stored as well as what sort of features should organizations now look for in backup software?*

Walter: I definitely see the same trend and I am in total agreement with you. Five years ago, much of our focus was on making sure that we had at least had one good copy of the data. Today, we seem to have sufficiently solved that problem. So with that being the case, I think now the time has come when we need to really focus more of our attention on the recovery piece of the puzzle.

Already, there's been a bit of a change in the backup infrastructure to accommodate for the introduction of disk into the backup process. But the next step, one that is coming on very strongly, is a keen focus on replication capabilities. Customers are finding out that quickly recovering virtual servers and large data sets – sometimes as quickly as within an hour – requires the use of frequent replication. That requires a different approach to data protection.

So we have reached a point where users can no longer just focus on backup jobs. Now, they need to focus on the needs of their organization in order to ensure that they can restore mission-critical applications within their respective recovery time objectives. That means thinking not just about data recovery, but operational recovery.

So, what do I mean by that? Not too long ago, backup or data protection mostly defined its scope around the ability to recover data. Now, not only do we need to focus on recovering the data, but we also need to focus on getting the email application or the transactional database back up and running. And fast.

That requires expanding the scope of backup and recovery outside of the data center, and it requires a shift from an infrastructure-centric view to a services-centric view that focuses on the actual services that IT needs to provide.

That's the big shift in focus here; the paradigm shift, if you will. Before, it made sense to focus on storage and on the infrastructure. But the shift that has to happen now is one that results in a new focus; a focus on operational recovery. It's going to center around our ability to enable fast application recoveries. So whether it's the ability to granularly recover critical items that have been lost, or to provide a fast way of doing a full application or database recovery, that's where we need to be headed as an industry.

Jerome: *That brings up an interesting point. Many organizations are seeing year over year data growth of 50% or more as we shift to capturing more information about people and from devices rather than people simply creating data which was the way it was traditionally created in the past. How does that change the challenges associated with the rapid recoveries of critical applications and databases in light of shrinking recovery times? Further, what needs to happen to meet the service level agreements (SLAs) associated with them?*

Walter: As we deal with explosive data growth and the need to quickly recover critical applications, we are going to have a strong need for a tiered approach to data protection and recovery. We'll see the continued adoption of snapshots, CDP, and replication to supplement full and block-level incremental backups, which are then deployed in targeted parts of the environment to make sure that IT can deliver on specific SLAs.

IT can no longer afford to have lower tier forms of data protection for the parts of the environment that have lots of data or require quick recoveries. The one size fits all approach to data protection will no longer work. Organizations will have to get more specific as to what types of data protection technologies they are going to deploy in their environment, and for what applications.

Part of the challenge with that is that most legacy backup products on the market today have a very infrastructure centric view of the world. Instead of seeing all of the data associated with applications, the infrastructure-centric view of the world relates all data to a specific server, and that makes it really hard for organizations to understand how a portion of their infrastructure services a certain part of the business.

So again, what we're going to have to do is change the way we view backup and recovery, such that we take a more services-centric view based on predefined SLAs. This will make it easier for organizations to use specific data protection tools for specific applications in order to achieve specific SLAs, while keeping other servers and applications on a lower cost data protection tier.

In [Part I](#) of this interview series Walter and I discuss how backup is changing and examine the quantum leaps forward that have occurred in how backup and recovery are done.

In [Part III](#) of this interview series, we look at how backup

software is evolving in light of the new challenges that server virtualization creates, in order to become smarter, more agile and do a lot more than backup.

In [Part IV](#) of this interview series, we explore whether or not virtualization only backup software solutions can survive long term.

In [Part V](#) of this interview series, Quest Software lays out its future plans for vRanger and NetVault Backup.

Amazon Fixes Dropbox Enterprise Gap; File Share and Synch Takes Off

Amazon [announced](#) their Storage Gateway (beta) on January 25th, about two days before my [article](#) on VMWare and Citrix squaring off in the “Dropbox for Enterprise” market. In my article I noted that VMWare and Citrix are exploiting an enterprise based limitation of Dropbox, Evernote and Box introduced by supporting a Consumerization of IT (CoIT) product. **Consumer-based file-share-and-synch applications** cannot be installed in a company’s data center. As file-share-synch drives cloud adoption in the enterprise, vendors emerge from all corners.

The Storage Gateway is initially targeted at an easy-to-deploy-in-the-cloud workload – **backup**. Moving backup data to the cloud requires the least amount contract negotiations between business units, IT and cloud service providers.

The Storage Gateway provides Amazon a **NEW** engagement with

Enterprises. Engaging with **data center managers** enables Amazon to develop products designed for on-premises deployments.

As the Storage Gateway product and deployment matures, Amazon's ecosystem of applications can leverage the on-premises footprint. Specifically, Dropbox, which has not had an on-premises version of file-share-and-synch, can work with enterprises to address that gap.

From the AWS Storage Gateway FAQ:

Q. How does the AWS Storage Gateway work?

*"...Data written to your Gateway-Stored volumes is **stored on your on-premises storage hardware, and asynchronously stored in Amazon S3 in the form of Amazon EBS snapshots**. This provides your on-premises applications with low-latency access to their entire data sets, while providing durable, off-site backups..."*

It is clear from the FAQ, that storing the data onsite is clearly in the plan. AWS S3 will be used as a backup for snapshots. This is akin to NetApp's [SnapMirror](#) functionality for snapshots.

(Note: Expect the low price "snapmirror like functionality" from Amazon to be a burden on both EMC (NASDAQ:EMC) and NetApp (NASDAQ:NTAP) high margin software offerings in the coming fiscal years.)

File-Share-and-Synch Vendors Emerge; what do we call this?

New vendors offering file-share-and-synchronization storage and application systems have approached DCIG in the last couple of months. However, each vendor has a slightly different name for the same market. In the midst of every conversation, I ask them – **what are you calling this market place?**

Vendor responses have varied from “*that’s your job as analyst*” to “*dropbox like solution for the enterprise.*” Buyers are struggling with what to call the market and products that service it. In the end, buyers suffer because it is difficult to identify products, competitors and overall requirements.

Further research of the market exposed “Mobile Collaboration” based on research Forrester was doing. Forrester’s assessment of the available products in 2011 for smartphones and tablets was accurate, it failed to capture the source of the data made available for collaboration in those devices – ***enterprise file and content management systems.***

Client availability in an App Store for a smartphone or tablet is a minimum barrier to entry for ***file-share-and-synch infrastructure.*** In addition, an on-premises installation is a crucial requirement for infrastructure as shared by vendors and enterprise buyers.

Vendors offering released products that address file-share-and-synch, mobile application and on-premises storage requirement include:

- [Accellion](#)
- [Citrix ShareFile](#)
- [GroupLogic MobileEcho](#)
- [ionGrid](#)
- [Mezeo](#)

Network Security Performance

Tuning by MetaFlows CEO Livio Ricciulli, Part II

Network security monitoring is a constantly changing environment of both tools and methodologies. Most of them today, however, have used a lone “cowboy” mentality where datacenter solutions operate independently. [MetaFlows](#) is changing that. Today, I am continuing my interview with MetaFlows CEO Livio Ricciulli, discussing how their product is optimizing network security monitoring and performance.

Joshua: *The MetaFlows product is delivering Network and Log monitoring facilities through a cloud-based “Software as a Service” model. I am sure one of our readers concerns will be that of performance. How does MetaFlows stack up compared to a solution where all the hardware and software is hosted within a company’s own network?*

Livio: We have placed a lot of effort in optimizing the performance of our software. In the past if you had a fairly fast network you had to buy appliances specifically designed for security, ranging in cost from \$20,000 to \$80,000 per gigabit per second. If you have multiple gigabits, they cost even more. These are specialized hardware devices that are highly optimized for security applications. So what MetaFlows did was to take our product and tune it for very high performance on off-the-shelf hardware. We reduced the cost of network security over that of deploying hardware appliances by using hardware that you can buy from any hardware vendor, and then creating a software library that parallelizes the processing.

For example, you can buy a \$1,000 machine with a nice set of Intel processors and process up to 800 megabits of data per second sustained with our software, which was unheard of before we added this capability.

Joshua: *I want to talk briefly about your competition. Who are your primary competitors? Is [Splunk](#) one of them?*

Livio: I think our innovation is aggregating a lot of functions into one system that have traditionally been split up between different vendors. *Splunk* does the log management portion. There are others doing the intrusion detection system function. Then there are still others doing flow monitoring, like [Arbor Networks](#). What we have done is to take pieces from all these different functions and aggregated them into one product giving companies the best of all these tools in one software suite. So I would say that *Splunk* is a competitor, but they are not a head on competitor.

Joshua: *So, then, your biggest competitor might be more operational issues within a company. That competition is convincing some organizations that using a disparate set of products is orthogonal to the operational success of their company. Using MetaFlows SaaS model, a company gets a best in class toolkit, and then does not have to worry about building installation, management and configuration expertise because MetaFlows is doing all that for them.*

Livio: Exactly. We have a conceptualization of the market today in four different solutions. There are low-end appliances, high-end appliances, open source solutions, and MetaFlows. The low-end is typically just an implementation of a particular open source IDS. It is low-end in the sense that it is not very sophisticated, and costs \$20,000. There is an initial 20 percent markup for a subscription to signature updates. To operate these appliances effectively, you need an expert on staff to interpret what is going on in that appliance and to interpret the output.

Next, there are the high-end appliances that are much more expensive. There is a higher subscription cost, but the administrator does not need to be an expert. You will pay around \$50,000 a year for them.

Then there is the open source route where you put a lot of time and effort into build something on your own. But you still need an expert to administer it because you need to be able to update and manage it yourself.

What we have done it to give you a solution at the cost of an open source product with a minimal amount of subscription cost, \$99 per month per CPU. You reduce the administrative costs because you do not need as many people to install and run it.

Joshua: *It seems like you can help companies better position themselves by having better tools and at the same time lower their costs. They will be able to deliver higher quality threat monitoring and threat identification while moving closer to cloud-enabling. Do you believe [Mobile First applications](#) and Cloud Storage should have a positive impact on the perception of your product, especially the global aggregation via Cloud?*

Livio: *Our take on this idea is that it takes a cloud to secure a cloud. The idea is that this architecture really is the best way to merge traditional hard-asset monitoring and cloud-based monitoring. Now, instead of having to host the database, you can disperse your agents globally, and have all of them point to one cloud-based system for storing events and logging everything that goes on in your network.*

[Last time](#), Ricciulli discussed how MetaFlows is delivering an innovative SaaS-based network security solution. Next time, he will explain how MetaFlows deals with payload security and how they deliver threat information to the end-user.

Network Security Monitoring delivered through a “Software as a Service” Model by MetaFlows CEO Livio Ricciulli, Part I

Enterprise organizations face the daily challenge of ever-growing threats to their network and IT infrastructure. Not only are these threats growing, but they are constantly changing as well, forcing companies to adapt by changing not only their tools but also their training. Today, I’m talking with MetaFlows CEO Livio Ricciulli about how [MetaFlows](#) is addressing these problems by delivering network security monitoring using the “[Software as a Service](#)” model.

Joshua: *Livio, thanks for taking the time to talk with me today. The network security monitoring space is filled with players today, but nobody is delivering a security product quite the way MetaFlows does. Can you talk to me a little bit about what makes your company and product unique in such a crowded space like network security monitoring?*

Livio: We’re the only system in the world that can do network security monitoring using the software as a service model. What that means is that the user does not host any of the system, except a very small portion of it – the agent that detects the threats – on their own network. The database where the events are stored, and all the software that analyzes the events, creates reports, and does the analysis, reside on a web server. You’re probably familiar with this model used by other “software as a service” applications.

Joshua: *Salesforce is a good example.*

Livio: Another example would be accounting with Quickbooks online – you don't host the data and the application. Everything goes through a browser. We're the first to provide a complete enterprise-ready, very sophisticated dashboard for monitoring the security of an enterprise entirely through a browser. Users simply install an agent that runs in the network and feeds events to a system that resides in the cloud. Then, the user can basically forget about the agent. It just runs there all the time. Everything that you need can then be found in the browser-based dashboard.

This promotes online collaboration between multiple analysts because you can access the data from a secure browser from anywhere in an environment in a way that's as secure as online banking.

Joshua: *How does this affect budget and existing analyst workload?*

Livio: This solution does two things. On the one hand it makes network security monitoring cheaper because most of the processing and the software updates are all taken care of centrally. When we do an update, it is immediately available to all our users. This is common to other software as a service services. But in the realm of security, it's an added value.

Another value-add is correlation. Our customers aggregate all their event data to a single location – our cloud. With that volume of data, MetaFlows can do a more effective job of correlating events across enterprises.

Joshua: *So are can you cleanse the data for analysts, such that they focus on critical threats across all data centers?*

Livio: We've also developed our own [algorithm](#) to rank security threats. This will help us to improve security as

well because it will alert people with a similar security posture to pay more attention to certain events than others. This type of targeting is not possible using a traditional model where everybody stores their own events in their own database and they don't share any information. This is very new. Nobody else is doing this as far as I know.

Joshua: *Can you reduce the amount of time security teams spend implementing, configuring and maintaining on-premise software?*

Livio: MetaFlows supports off-the-shelf hardware, meaning you can download our agents for practically any hardware you can buy. We also sell inexpensive appliances for those wanting a more traditional hardware-based solution. In either case, you download an agent from our website that then does traffic monitoring in the enterprise. It includes a suite of applications that are designed to give very broad detection capabilities ranging from looking for [bots](#) – computers that are subverted to become Trojan horses- to more of a generic [intrusion detection system](#), where we look for events, like somebody using peer to peer file sharing, that is not a permitted use of the network.

The agent does the monitoring, and when an event is detected, the event data, not the payload just the event, gets fed to our MetaFlows cloud in real time. These events then get stored, correlated, and archived in the cloud. Then, users can interact with Metaflows to look at events, correlate them, and get a picture of what is going on in their network, all through a browser.

Joshua: *Since this is security log data, can you include other log data from devices on a network?*

Livio: We also support log management. We can store all the events logged by third-party devices on the network. So, essentially, we can unify the IDS function and the log

management function all in one place. And this solution is turn-key, you don't have to install, configure and setup anything, just download it and run it.

In the [next](#) blog entry in this interview series with Ricciulli he will explain how MetaFlows is optimizing network security monitoring and performance.