

Coping with “Big Data” Growing Pains

Lis Strenger



Lis Strenger is the director of product marketing at Dataupia.
lstrenger@dataupia.com

Abstract

Data volumes are growing exponentially. That’s a given. How organizations respond to the growth, however, is not. A recent survey of data center managers shows that newer practices such as single-instance storage are gaining traction but not replacing older methods.

To stay on top of data growth, IT groups are combining approaches such as data deduplication, archiving, deletion, workload prioritization, hardware acquisition, hosted repositories, and database optimization. However, most are still making decisions about handling “big data” based on their previous experience with “small data,” and risk making costly choices or missing opportunities to improve information management.

This article explains how “big data” changes the rules of the game. Strategies that worked well for gigabytes of data fall short in the terabyte world. We answer the top five questions data warehousing professionals have about working with terabytes. After looking at common industry approaches under the terabyte lens, readers will have a relevant framework for evaluating technology solutions and business process changes to support their growing data assets.

Introduction

I recently read a short report in *Time* about a Russian fast-food chain that has swiftly grown to become fourth largest in that nation (McGrane, 2008). The founder, Mikhail Goncharov, made a comment about managing growth that caught my attention: “If you’re chopping 100 kg of mushrooms, you do it one way. If it’s 200 kg of mushrooms, you do it a totally different way.” That, in a nutshell, is the mindset to have when facing the need to manage large data volumes.

How do you find that different way to chop twice the amount of mushrooms? Do you pull out the colanders and knives you used yesterday and see how far they get you? Of course, you count on your prep-chef experience to tell you when and how to make adjustments along the way. Do you reach for the largest cleaver you have, following the principle of “the larger the volume, the larger the tool,” or do you look at the tower of boxes and rethink the whole procedure? The latter choice isn’t likely if you are running a real business with limited resources and concrete requirements—although you might start researching appliances designed specifically to chop fungi.

Instead of counting terabytes or the number of data rows you store, determine whether the amount of data you need for operations exceeds the capacity of a component within your infrastructure by a few gigabytes.

Begin to think through how to handle large amounts of data by evaluating your current tools and deciding what you can scale (what can be extended) with supporting technology. Perhaps in parallel, research other available solutions, mindful that it is unlikely you will find the perfect fit. You will have to tailor a solution to meet your requirements. Whatever approach you emphasize when looking for solutions, there are two things you need to know when you develop your requirements and evaluate options:

- Define “big data” in your organization’s or project’s context. Technology vendors, industry analysts, and academics each have a different definition. However, as well-founded and valid as these definitions are, they will not apply directly to your situation. Understand

where your data comes from, the business context of each data set, and where data volume is largest.

- Know the capacity and extensibility of the tools and methods you are investigating, especially with respect to how they perform at the top range of your forecast. *Predictions about how our information environments will expand are usually accurate.*

Taking Measure of Your Data

Data complexity is a result of working within memory, disk space, CPU, I/O, and network constraints. An operation executed against large data sets is split into subtransactions or subprocesses, and continuity must be guaranteed. That threshold is shifting as advances in technology ease some of the constraints, but there is still a size barrier. Like the sound barrier, the size barrier can be crossed, but some of the rules change on the other side.

There is a more pragmatic way to determine whether your data qualifies as “big.” Instead of counting terabytes or the number of data rows you store, determine whether the amount of data you need for operations exceeds the capacity of a component within your infrastructure by a few gigabytes (GB). If so, you face many of the same scalability challenges as someone trying to add terabytes (TB) of data. For example, Microsoft SQL Server has its own internal size markers. Its COUNT function tops out at two billion records, after which COUNT BIG must be used. This is just one example of how, at the fundamental level of query processing, size changes the rules. COUNT BIG mimics COUNT as much as possible, but the differences can have a ripple effect. The database can operate against a big data set, but can your infrastructure support passing a big result set through the application stack?

Your data warehouse is swelling. You know it takes in 500 GB per day, but from how many feeds and at what rate? To understand the impact big data is having on your infrastructure, identify the points in your data management infrastructure that touch large volumes of data, either serially or all at once. These parameters are critical when architecting your solution. Managing two 200 GB batches every 12 hours is different from managing multiple lightweight sources fed hourly.

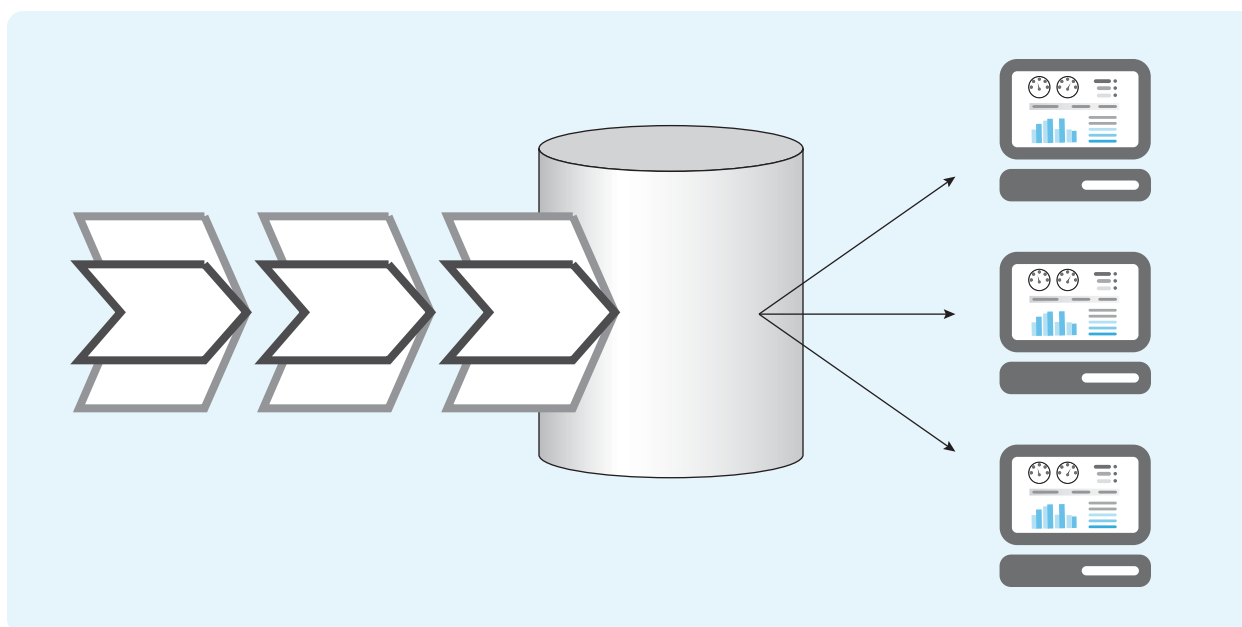


Figure 1: Some data warehouses must scale to adapt to the growing size of data sets. In this illustration, only the size of batch feeds grows, not the frequency or the degree of application access.

Figures 1 and 2 show examples of the factors that require a data warehouse to scale. Although these examples look simple, accommodating just one growth area can result in work at almost every layer of the data warehouse's infrastructure and surrounding networks and applications. The type of changes required by growth factors will vary, however. It is essential to map all the ways a data warehouse will grow at the outset of any expansion project. Consider whether it will need to accommodate larger data sets, more data sets, higher frequency of data loads, more users, more applications, and so on.

Assessing Impact Points and Approaches

Just as there are many ways a data warehouse might need to grow, there are many points at which its infrastructure must scale. Figure 3 (page 49) shows the many points where opportunities exist to make big data more manageable: at the point of collection, before sending data to the data warehouse, before archiving, and before off-site storage. Each is associated with performance or accessibility trade-offs. For example, the more that is done at the point of data collection, the more performance degrades from the users' perspective. Shifting all work to the end of the process puts pressure on maintenance

procedures; backup and replication might collide with data collection and access.

Figure 3 shows a solution map you can use to identify your anticipated growth areas. This is the first step to selecting approaches that address the particular challenges your data warehouse faces.

The second step in the process is to look at the short- and long-term impact each approach has on both the IT and business users. To measure "impact," examine expense, labor, time, future scalability, and whether accessibility to data increases or decreases for the business user. Table 1 (page 50) compares the relative impact of each approach or technology.

The art in applying this information lies in striking a balance among all factors that your organization can live with. No single practice stands out as being a panacea. Instead, your organization's tolerance for limited accessibility to data, your IT group's appetite for taking on new technology, and resource constraints will determine a solution's viability.

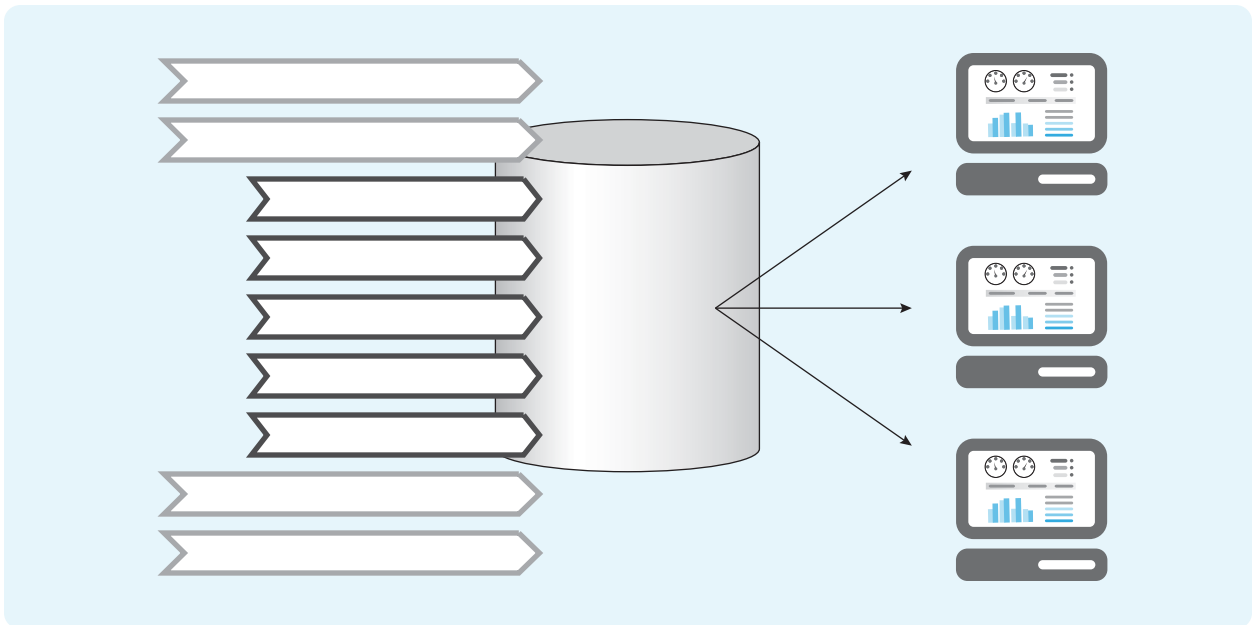


Figure 2: Adding new data sets, even small ones, can cause scalability challenges for a data warehouse even if all other parts of the environment stay the same.

Up to this point you've identified the sources of your growth and the stress points the added volume will cause in your data warehouse infrastructure. You've been introduced to the approaches that ease scalability problems, identified where in the architecture they operate, and determined what factors will play into your choices. That's the framework for evaluating solutions. You still need information to evaluate the solutions on their own merit.

Use Existing Capacity More Effectively

The clear first task is to ensure that existing resources are used appropriately and to their fullest extent. Techniques that help stretch existing resources include work prioritization, system optimization, and virtualization. However, these approaches can provide only a temporary fix, because your existing resources were selected to fulfill very different requirements (at least in terms of scale) than those you have now.

Using existing capacity to its fullest is a matter of applying basic resource allocation techniques, although now you are making rules based on a new set of conditions. You still have two options: to manage

hardware or manage users. When classifying units of work for resource allocation or for query prioritization, size becomes critical. In particular, you are no longer concerned only with the number of CPU cycles a request consumes. You must also consider the time it takes to access secondary storage and the application impact of moving several terabytes of data across the network.

Assessing the impact of an operation requires an in-depth knowledge of the database, its table sizes, data density, and indexing. Scheduling with an eye to minimizing resource conflicts is a possible solution, though low-usage windows are scarce in multinational organizations. The result is that prioritization will increase your actual capacity but at the cost of additional complexity, expertise, and slower system performance.

Established practices of managing workload using query prioritization and job scheduling are still effective here, though you will have to revisit the algorithms you use and the business rules they embody. For example, standard reports, when run against massive data sets, can consume more system bandwidth than the business will tolerate. Larger data sets also present new opportunities

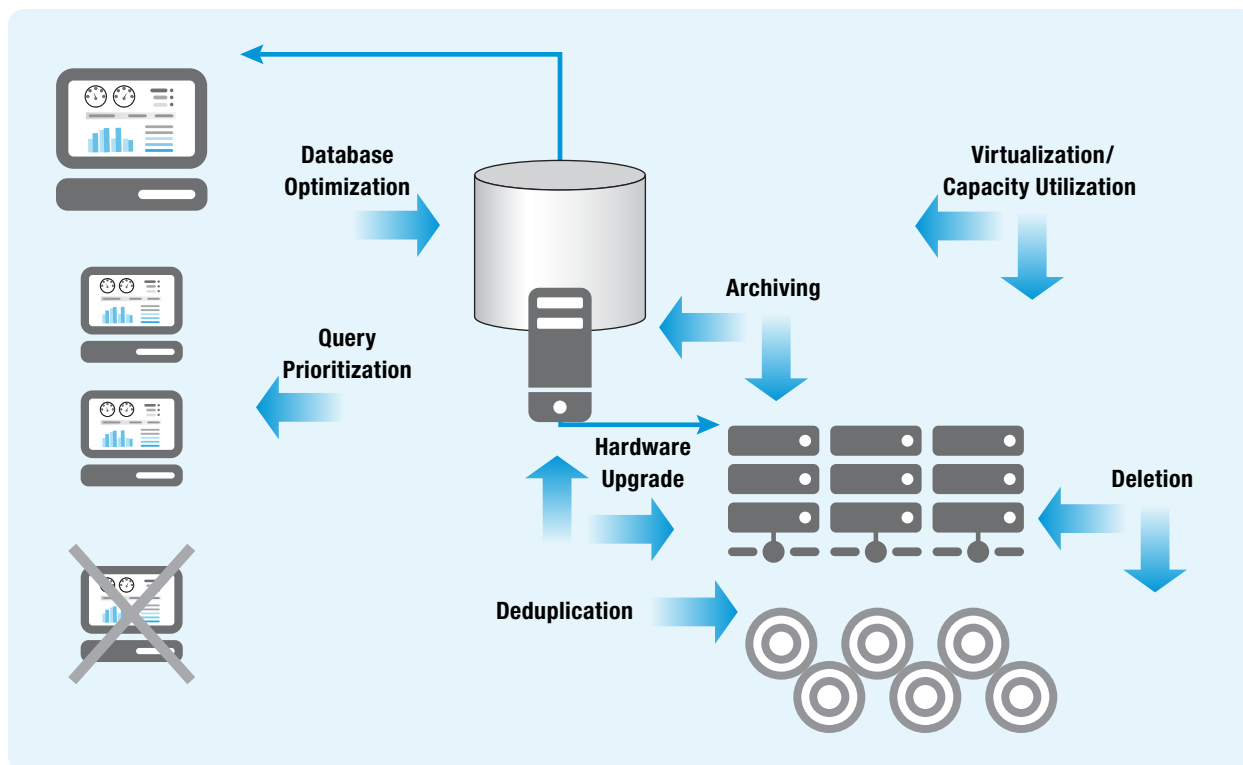


Figure 3: The approaches to handling a data warehouse's growth increase scalability at specific points in the architecture.

to gather information, especially through analysis. The standard reports once viewed as mission-critical may be less desirable if running them limits root-cause or trend analysis that can provide more meaningful information.

Optimize Databases

When you optimize your database for large volumes, remember that the database platform you are using was designed for OLTP, not for data warehousing. Although you might have built your data warehouse from the ground up, the technology you used had many built-in assumptions stemming from its OLTP roots. If you transformed an OLTP database into a data warehouse, you should review for transactional legacy even more rigorously. For example, are cursor numbers, buffer sizes, and limiting parameters set appropriately? Have you fully exploited your database platform's native optimization capabilities, such as partitioning, distribution, and query tuning?

Deploy Virtualization

A 2008 *McKinsey Quarterly* article describes a common problem businesses have with unused capacity:

Well-managed companies use 80 percent or more of their available storage, but in others that figure hovers around 40 to 50 percent. One large IT organization used only 50 percent of its storage capabilities. Some of its individual storage systems were at just 10 to 20 percent of capacity, and one of its businesses utilized only 33 percent of the entire amount of storage it had requested.

Technologies such as virtualization have a strong impact on the ability to tap into a broad range of existing resources as needed. Virtualization alone cannot address the challenges of working with large volumes of data, but it can stretch your resources, especially where you have many small streams of data or a high rate of data operations. If your large data is made up of smaller data

	HARDWARE COST	LABOR	TIME	SCALABILITY	LIMITED ACCESSIBILITY
DATABASE OPTIMIZATION	Low	Moderate	Low	Moderate	High
HARDWARE UPGRADES	High	High	Moderate	High	None
VIRTUALIZATION	Moderate	Moderate	Low	Moderate	Low
DATA DEDUPLICATION	Moderate	Moderate	High	Moderate	Low
ARCHIVING	Low	High	Low	Low	High
DELETION	None	Low	Low	Low	High

Table 1: Impacts of each approach to, or technology for, handling growth

sets, you can farm out the units of work to available CPUs and distribute the data blocks to disk space spread through the virtual array. However, if you intend to run operations against the whole data set, neither virtualized servers nor virtualized storage will suffice. At some point, you will need enough CPU and disk capacity that can function as a single engine. Virtualization management tools are not yet advanced enough to distribute a single unit of work and guarantee its integrity.

Acquire or Upgrade Hardware

Virtualization will extend some of the capabilities of your physical infrastructure until data growth exceeds your hardware’s capacity. More often than not, adding or upgrading hardware is the first solution data center managers turn to when confronted with growing data. The first signs of problems show up as hardware performance issues signaled by high CPU activity, hanging processes, and running out of memory or disk. Adding more hardware without altering the physical architecture will not address the root cause, which is the need to move too much data through the network.

For physical infrastructure to support working with big data, three components must be scalable: CPUs, disk I/O, and network connectivity. The latter is important because unless you have a mainframe, your big data cannot be processed and reside on the same server. A typical architecture for big data has at its core a robust, multi-CPU server hosting the database and network-attached storage. There might be additional servers for auditing services, backup facilities, ETL, data cleansing, data staging, etc. You can add more CPUs at the front end and you can add more storage,

but you also need to address bandwidth by moving to more powerful backbones (Fibre Channel or 10 Gigabit Ethernet) and switches.

Databases that have tens or hundreds of terabytes require some form of parallel processing. Massively parallel processing (MPP) is becoming the best-in-class architecture for very large databases. MPP configurations are designed to recast operations as sets of subprocesses, distribute them for parallel execution on an array of CPUs and disks, and marshal the results. Both the hardware and the database have to be designed for MPP for this level of parallelism to occur. IBM’s DB2 was an early MPP database, but most deployments are on servers that have minimal parallel processing. MPP systems are available as data warehouse appliances or as specially configured hardware/software bundles.

Adjust Information Lifecycle Management Policies

Why are so many organizations overwhelmed by exponential increases in data? One of the main drivers behind this growth is that organizations have explicitly decided to collect these volumes of data. More demand has led to more data—more data will lead to increased demand.

Putting draconian archiving or purging policies in place would stem the data tide, but the goal is to maintain access to data despite size challenges. Information lifecycle management (ILM) based solely on age, access frequency, or compliance conflicts with the purpose of amassing data, namely to maintain enough historical, detailed data to support strategic and tactical decision making. Even data that is collected and retained primarily to satisfy regulations can be mined or analyzed for

trends and patterns. A recent study by Nemertes Research on security and information protection showed that 27 percent of participants kept compliance data “forever.” Leveraging that data for intelligence gathering would make up for some of the cost of housing it (Burke, 2008).

To the extent that big data is putting a strain on resources, you will want to refine the processes by which you rank projects, subject areas, and data sets. Age of data takes on a different meaning if the business has decided it wants to collect clickstream or shopping cart data for three years to discover seasonal customer behavior. Operational data like this would once have just been analyzed on the fly, then deleted. Now it becomes more significant and has a longer shelf life. Although it won’t be accessed often, analysts and management may want it on demand, as might customer service reps in companies embracing operational BI.

Even compliance to data’s ILM might have to change. Since the costs of compliance are so high, the business might want to recoup some of that investment by leveraging it for its BI value, which means that it needs to stay online longer instead of being archived to less expensive storage as quickly as possible. One solution is to treat read-only data differently from data that continues to be updated. Instead of two stages (online and archive), you would use three stages: online current, online read-only, and archive. Read-only data requires fewer resources. Creating and maintaining a separate repository for rolling off read-only data is not a trivial task, but reducing the pressure on resources might justify this additional step.

These changes in data usage have more to do with the reasons organizations are collecting so much more data than with the issue of big data. Negotiations on service-level agreements will have to balance the increased demand against resource availability. New ILM guidelines need to take business significance into account. ILM decisions have typically been made primarily in the data center in consultation with data governance and risk management groups who represented the business. Now, sponsors of the many input data streams will be needed to provide the business context for ILM and resource allocation decisions.

Research Deduplication—An Emerging Practice

A single terabyte of data needs 53 terabytes of storage over its lifetime because of the number of times it is instantiated across multiple applications or data marts, time series snapshots, backups, and replication (Darrow, 2008). Deduplication techniques reduce these many images to a single one, which promises to decrease storage needs by a factor of 20. As a result, 16 percent of data center managers surveyed by InfoPro Incorporated plan to adopt deduplication within a year. For the first time in several years, entrenched data center practices are losing ground to a relatively new technology.

To the extent that big data is putting a strain on resources, you will want to refine the processes by which you rank projects, subject areas, and data sets. Age of data takes on a different meaning if the business has decided it wants to collect clickstream or shopping cart data for three years to discover seasonal customer behavior.

Business users reading this description might jump to the conclusion that deduplication will bring about the elusive “single version of the truth,” but the impact of deduplication is far removed from the business user. The single image refers to a single image of data on backup tapes or in third-tier storage. As such it is also removed from data integration and master data management initiatives. As the technology evolves, we might see it deployed as part of the data warehouse.

Although deduplication will not reduce the amount of data in a data warehouse, it will reduce the number of

tapes or disks in your physical warehouse and lower end-of-life storage costs. Deduping technology is available via software and hardware solutions. It can be implemented off-line (that is, after data has been stored) or in-line (as the last step before data is archived). Even this technology, which seems tailor-made for big-data scenarios, has hard limitations when it comes to disk space. Most solutions require that you first store the data on a storage server, then run it through the deduping procedure. Today's solutions cannot accommodate more than 20 TB for off-line deduping.

A deduping implementation requires the same amount of storage up front—a 20 TB database needs 20 TB of storage space. The reduction comes once a history of deduping the same database is established. Applying the 1-to-53 rule, the initial TB will require 1 TB, but 53 subsequent snapshots will not require significantly more space than that initial TB. The advantage to such reduction is clear: backup and replication will be more efficient, hardware costs will be reduced, and energy and space requirements will be minimized.

Design Your Solution

Retrofitting your infrastructure to handle massive amounts of data is a complicated proposal. There is no single tool or methodology for scaling resources to keep pace with data growth. Neither is there a set of tools or a reference blueprint that lays out a clear path. Best practices are just now beginning to emerge as more organizations cross the size barrier. Even today, on community sites dedicated to SQL, database analysts post questions about how peers handle environments for databases larger than 100 GB.

Most of us are taking inventory of current resources—hardware, software, and expertise—and seeing how far they'll take us. At some point, however, we have to shift perspective, because the size of data changes many of our assumptions about collecting, using, and storing data. The methods that have made data centers more efficient and data more secure need careful review to uncover the weaknesses that appear only when working with large amounts of data. Even the business and usage policies

have to be adjusted to scale along with the data and to accommodate new usage patterns.

Even if there is no prescribed approach to managing data growth, there are some guidelines for evaluating solutions and whether they fit your own environment and requirements. Any operation on large data will depend on CPU capacity, network bandwidth, and disk I/O. Many expansion initiatives don't show the expected results because of a failure to consider all three of these components. The burden on these components can be addressed through database and query optimization, workload prioritization, virtualization, hardware investments, or usage and ILM policies.

New data management approaches must address the sheer presence of so much data as well as anticipate how the users change what they expect from their big data and how they leverage it. In other words, don't let the storage dimension overshadow access considerations. The world of compliance offers up some insight here. Let's not forget that every data retention regulation has a "timely access" clause. ■

References

- Burke, John [2008]. "Compliance-Related Costs are Rising," *NetworkWorld Executive Guide: Storage Heats Up*, white paper, page 18.
- Darrow, Barbara [2008]. "Is Fibre Channel Dead?" *NetworkWorld Executive Guide: Storage Heats Up*, white paper, page 5.
- McGrane, Sally [2008]. "The Czar of Crepes," *Time*, June 5. <http://www.time.com/time/magazine/article/0,9171,1812074,00.html>
- The McKinsey Quarterly* [2008]. "Meeting the Demand for Data Storage," June. http://www.mckinseyquarterly.com/Information_Technology/Management/Meeting_the_demand_for_data_storage_2153?gp=1