The Challenges and Rewards of Big Data

an INFOSTOR® Storage eBook

Contents...

The Challenges and Rewards of Big Data

This content was originally published on the IT Business Edge, Enterprise Apps Today and Enterprise Storage Forum websites. Contributors: Loraine Lawson, Pedro Hernandez, Arthur Cole, Vangie Beal and Henry Newman.

What's the Big Deal About Big Data?

By Loraine Lawson

Big Data is big news these days. Still, I'm sure there are those among you wondering whether Big Data is actually a big deal or just a big bloated bag of hot air.

It's a legitimate question, and you're not alone in asking it. In fact, McKinsey Global Institute (MGI) and McKinsey Business Technology Office studied the issue in-depth in the report, "Big Data: The Next Frontier for Innovation, Competition and Productivity." The conclusion: Yes, it is a big deal, and it has huge potential in every sector examined.

The study examined five areas:

- U.S. health care
- U.S. retailers
- Europe's public sector
- Worldwide manufacturing
- Global personal location data

The 156-page report is available for free download and includes a list of ways Big Data can create new opportunities and savings:

- $300 billion more value each year for the U.S. health care system, two-thirds of which would come in reduced expenditures
- Up to a 50 percent decrease in product development and assembly costs for manufacturing
- $149 billion worth of operational efficiencies for European governments
- A potential for retailers to grow operating margin by 60 percent

McKinsey not only found the potential for huge benefits from using Big Data, it also found early adopters putting it to use in every sector.

What's more, there is enough Big Data accumulated to put it to use today. The report found:

> The growth of big data is a phenomenon that we have observed in every sector. More important, data intensity — i.e., the average amount of data stored per company — across sectors in the global economy is sufficient for companies to use techniques enabled by large datasets to drive value (although some sectors had significantly higher data intensity than others).

McKinsey looked at whether the average user would benefit and discovered $600 billion worth of savings to consumers based on personal location data alone. For example, travelers would shave 10 to 15 hours off drive time, equaling an annual savings of roughly $150 billion in fuel consumption, according to a report summary published on the Financial Times (free registration

required). On behalf of average users everywhere, I'd just like to say bless 'em for that news.

I won't go into Big Data's privacy cost, and neither does McKinsey. Suffice it to say I think Big Data makes Facebook, Apple and GPS look like mere spy toys, and privacy will be as obsolete as Kodachrome film for everyone but the Amish.

But as you might expect, there are challenges, particularly around integration.

Big Data's Integration Hurdles

Big Data is not without its complications, and integration is chief among them. Let's look at some integration issues you'll need to consider when it comes to Big Data.

First, it's important to note that when the McKinsey analysts and others talk about Big Data, they're really not just talking about massive amounts of data, as Forrester Principal Analyst Brian Hopkins points out:

> We are thinking about it in terms of not only big volume, but high velocity, variety and variability. Some of the most interesting uses of technologies such as Hadoop are coming from the velocity and variability characteristics of "data at an extreme scale" — which is perhaps a better thing to think of when you hear the words Big Data.' What we are seeing is that it's not about just handling large amounts of complex data — agree, we have been doing that for years, as you point out. It's more about handling it in ways that are faster, cheaper and more forward looking that our current technology allows.

The McKinsey analysts take essentially the same view, as they explain in the Financial Times article:

> In addition to the sheer scale of big data, the real-time and high frequency nature of the data is also key. For example, 'nowcasting' is used extensively and adds considerable power to prediction. Similarly the high frequency of data allows users to test theories in near real-time and to a level never before possible.

Obviously, this focus on real time and high frequency requires a more robust integration plan than your run-of-the-mill extract, transform, and load (ETL) project.

Not that ETL is useless. Actually, it makes the list of "Big Data Technologies" included in the McKinsey report, starting on page 31. A number of commonplace technology items are listed, including metadata, mashups, data marts, data warehouses and cloud computing, as well as the names typically associated with Big Data, such as Hadoop, MapReduce, R and Cassandra. Big Data brings together a veritable "what's what" of integration and data management tools.

Again, many of the integration technology issues associated with Big Data are also not unique — they're just required on a uniquely large, fast and frequent scale. Legacy systems, and incompatible standards and formats are among the data integration challenges you'll encounter with Big Data. Again, that's nothing new.

But Big Data will require integrating a broader range of data, according to the Financial Times article. This is not your typical B2B fare:

> Above all, access to data needs to broaden. Increasingly companies will need to access data from third parties and integrate them with their own, but today there are few areas where there are efficient markets for the trading or sharing of data — for example, there is no market for the sharing of the aggregate movement patterns derived from mobile phones that retailers want to mine as they try to understand the behavior of their customers.

In many cases, we're talking about large data sets owned by governments, some of which aren't even available yet.

Obviously, Big Data will be a big deal, creating not just new business and cost-saving opportunities, but also potentially redefining democracy and open government. However, all of that potential hinges on our ability to integrate, mine and use the data. McKinsey's report offers a good starting point for turning Big Data's potential into a reality.

Big Data Faces Big Disconnect

By Pedro Hernandez

Big Data may be dominating the conversation in IT circles, but most businesses just can't seem to get on board yet.

While conducting a survey of 339 data management pros, SAS and SourceMedia discovered that for most, Big Data simply doesn't factor into the day-to-day operations of their organizations. Only 12 percent of respondents said they had a Big Data strategy currently in place.

The rest, suggests Todd Wright, global product marketing manager for SAS DataFlux Data Quality, are leaving money on the table. "The 12 percent of organizations that are already planning around big data enjoy a significant competitive advantage," he said in a company statement.

The study is the latest to highlight the seemingly uneasy relationship businesses have with Big Data solutions providers.

Seventy percent of organizations polled by integration software specialist Informatica told the company they were planning or had already pulled the trigger on Big Data projects. The survey also revealed that 71 percent of companies were looking to improve efficiency, and 50 percent sought to launch new products and services.

However, despite these aims, most businesses are struggling to leverage Big Data in meaningful ways.

Neolane, a maker of marketing software, and the Direct Marketers Association recently revealed that 60 percent of marketers don't have a handle on dealing with Big Data challenges. CompTIA, meanwhile, released a study that indicates the industry needs to do a better job of communicating what exactly Big Data is. Just 37 percent of IT and executives polled by the organizations said they were very familiar or mostly familiar with the concept of Big Data.

Those themes were echoed by the SAS study.

Twenty-one percent of those surveyed said they didn't know enough about Big Data, and 15 percent didn't understand its benefits. Business support was nonexistent for 9 percent of the survey takers, while another 9 percent found data quality lacking in existing systems.

These shortcomings, if left to linger, could impact providers of Big Data services. Only 14 percent said they were "very likely" to use external Big Data sources in 2014. A small but significant minority (19 percent) said they were "not likely at all" to incorporate external Big Data.

Companies also appear to be struggling to identify who takes the reins of Big Data within their organizations, sowing dysfunction.

"The survey found no real consensus on who owns the data management strategy, with responses ranging from midlevel IT personnel up to the CEO. This confusion likely causes additional challenges in data strategy development and execution," said SAS.

Although the picture looks grim, there is some evidence that businesses at least have a grasp of what they want from Big Data solutions. Seventy-three percent said they were seeking data visualizations and dashboards, while 53 percent had their sights set on data profiling. And there was good news for Big Data cloud specialists, 44 percent said they were looking for software-as-a-service (SaaS) offerings. ■

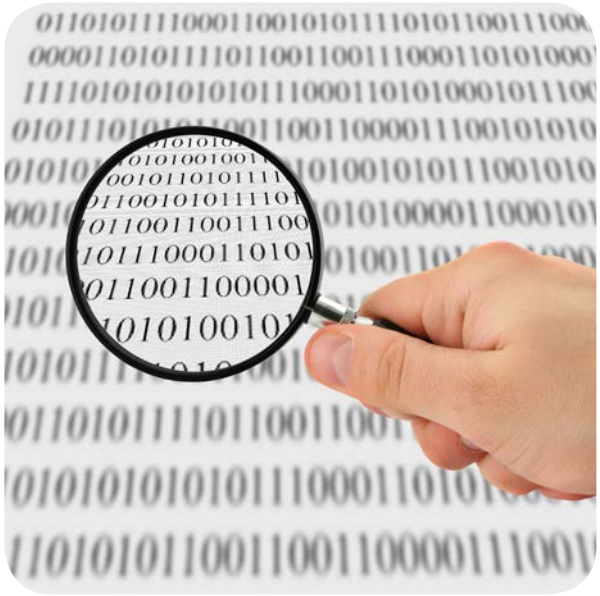
"Although the picture looks grim, there is some evidence that businesses at least have a grasp of what they want from Big Data solutions."

Big Data's Need for Big Solutions

By Arthur Cole

Big Data may be a fact of life for many enterprises, but that doesn't mean we are all fated to drown under giant waves of unintelligible and incomprehensible information.

In fact, the very definition of Big Data holds that loads can acquire the term only if they exceed the organization's ability to analyze and manage them. Therefore, Big Data is determined not by what is coming at you, but by your ability to handle it.



The answer, then, is to increase your ability to handle it. However, the speed at which Big Data is mounting and the complexity in building a suitable infrastructure add both cost and configuration hurdles to any Big Data program. At the same time, data is hitting the enterprise from a plethora of sources, particularly as new generations of mobile and handheld devices produce a deluge of unstructured data that defies easy analysis.

But before you think this is just a storage and analytics challenge, consider the impact this is having on the network. As WAN optimization provider Infineta Systems showed recently, data center-to-data center connectivity is likely to suffer the most in Big Data environments. As enterprises adopt Hadoop and other tools to increase capacity to the petabyte level, reliance on wide-area networks will increase as disparate resources are drawn together to process the load. Most current WAN environments are not tailored for extremely large data sets, leading to bottlenecks and performance deterioration.

Elsewhere in the stack, Big Data presents more of an opportunity than a challenge in light of changing data center platform requirements. Red Hat, for example, sees a chance to expand its footprint in enterprise environments through large-volume storage enhancements. The company's Storage Software Appliance leverages technology acquired from Gluster to provide scale into the petabyte range. As an open source, distributed file system, the system provides for virtualized, pooled storage that can be layered across physical and cloud environments.

At the same time, Big Data presents new opportunities for some long-standing enterprise solutions. As HP's Srinivasan Rajan points out, COBOL offers a number of advantages when it comes to handling large data sets. For one, Big Data analytics can take advantage of COBOL's batch infrastructure support and complex algorithms. As well, COBOL's Job Control Language (JCL) is highly adept at scheduling large jobs into smaller ones, helping to lessen the impact on underlying resources.

Big Data, it would seem, is neither a net negative nor a net positive for the enterprise. Yes, it presents a challenge both for analysis and physical/virtual resource allocation. But at the same time, there is likely to be tremendous value in all that noise. The trick will be to identify and leverage the crucial bits of information without overwhelming your data infrastructure or IT budget. ■

CIOs Invest in Business Analytics to Tackle Big Data

By Vangie Beal

According to an IBM global study of CIOs, the top strategic technology investment over the next five years at outperforming midsize organizations is using business analytics to extract actionable insights from Big Data.

The study also indicates that cloud computing has emerged as the fastest growing technology area for CIOs.

The IBM study looks at what constitutes the fundamental tasks of the CIO and what traits define outperforming CIOs as they infuse technology into products, services and processes to transform their businesses, drive profitability and expand into new areas.

The Essential CIO -- Midmarket CIO Study reflects the increasingly important role played by CIOs of midsize companies as the global economy continues to recover.

"As the economy recovers and CIOs look more to driving transformation in their companies, their role is evolving to become more and more associated with extracting value from technology and gaining insight from complex systems," said Ed Abrams, vice president of marketing for IBM's global midmarket business. "CIOs who consider themselves over-performers are aware of their increased impact on the organization and are steadily turning to analytics and cloud to pursue their visionary goals."

Big Data Identified by CIOs as Top Priority Investment

CIOs surveyed identified analytics, the ability to extract actionable insights from "Big Data," as their top-priority investment area by 83 percent. CIOs are looking to invest in technologies, such as analytics and data mining that not only help them better utilize structured data, but also use unstructured data in the form of videos, blogs and tweets that can be obtained through the social web.

Like their peers in larger companies, midmarket CIOs are facing an increasingly complex business environment defined by sweeping changes and the need for gaining greater intelligence, insight and visibility. These CIOs increasingly view tackling Big Data as a key imperative for gaining insight and expanding relationships with customers and partners.

The Shift to Mobile and Cloud Computing

There was also a 50 percent increase in the number of midsize organizations that plan to invest in cloud computing when compared to IBM's previous midmarket CIO study. CIOs are now 50 percent more likely to pursue investments in cloud over the next three to five years to take advantage of the flexibility and cost-effectiveness of using hardware and software resources offered through the cloud.

Trends such as the growth of Internet-connected devices and smartphones are driving CIOs to consider more powerful ways to harness mobile applications that drive commerce, better collaboration and enhanced workforce mobility.

According to the IBM study, the percentage of CIOs who plan to invest in mobility solutions increased by 11 percent over 2009 to 72 percent.

Disaster Recovery Planning for Large Archives

By Henry Newman

Disaster recovery (DR) is often discussed in broad terms throughout the storage industry, but in this article I will explore a specific segment of the overall market: DR planning for large archives.

What are my definitions of an archive, and what is a large archive? An archive is a repository of information that is saved, but most of the information is infrequently accessed.

The definitions of archives have changed recently. Just three or four years ago, archives were always on tape, with only a small disk cache (usually less than 5 percent of the total capacity). The software to manage data on tape or disk is called hierarchical storage management (HSM) and was developed for mainframes more than 35 years ago.

Today, we have large disk-based archives that back up data over networks. For example, both my work PC and home PCs are backed up via the Internet, and large cloud-based archives are common. There is, of course, a question of reliability, but that is a different topic.

My definition of a large archive is fairly simple: anything over 2,000 SATA disk drives. Today, that is about 4 PB, and next year, when drive capacities increase, it will likely be 8 PB. I am using 2,000 drives for the archive size given the expected failure rate of the 2,000 drives. Even in a RAID-6 configuration, which would require 2,400 drives, it will be challenging given the rebuild time to manage that many drives for a single application.

Three Types of Disaster

There are three types of disasters to be considered: failure of a single file or group of files, metadata corruption, and what I often call the "sprinkler error."

The failure of a single file or group of files is a completely different problem than a sprinkler going off in a computer room and destroying all of the equipment. The failure of a file or groups of files is significantly more likely and far more common than a complete disaster (e.g., earthquake, hurricane, lighting strike, power surge or sprinklers going off), but when I architect systems I ensure that there are always at least two copies of the data. In large archives, given the time to re-replicate the data and the data integrity from the storage system in the event of a disaster, two copies may not be enough.

The metadata corruption problem is also unlikely, but it does happen and it happens more often than many believe. Metadata corruption could be the corruption of the file system metadata or, if data deduplication is used, the corruption of one of the blocks which, if it is not well protected, will be a disaster.

Of course, cost plays a big part in how much data protection a site will have. Many vendors talk about four

9s, five 9s, or even eight 9s of availability and reliability. However, when you have many petabytes of data this concept must be re-thought.

The chart below shows expected data loss based on the number of 9s of reliability. Data loss is in Bytes.

Data Loss in Bytes

9s	Data Reliability %	1 PB	50 PB	100 PB	500 PB
2	99%	90,071,992,547,410	4,503,599,627,370,500	9,007,199,254,741,000	45,035,996,273,705,000
3	99.9%	9,007,199,254,741	450,359,962,737,050	900,719,925,474,100	4,503,599,627,370,500
4	99.99%	900,719,925,474	45,035,996,273,700	90,071,992,547,400	450,359,962,737,000
5	99.999%	90,071,992,547	4,503,599,627,350	9,007,199,254,700	45,035,996,273,500
6	99.9999%	9,007,199,255	450,359,962,750	900,719,925,500	4,503,599,627,500
7	99.99999%	900,719,925	45,035,996,250	90,071,992,500	450,359,962,500
8	99.999999%	90,071,993	4,503,599,650	9,007,199,300	45,035,996,500
9	99.9999999%	9,007,199	450,359,950	900,719,900	4,503,599,500
10	99.99999999%	900,720	45,036,000	90,072,000	450,360,000
15	99.99999999999999%	9	450	900	4,500
20	99.99999999999999999%	0	0	0	0

So for ten 9s of data reliability and just a single petabyte of data, a loss of 900,720 bytes can be expected. Therefore, data reliability in terms of a count of 9s must be reconsidered in the context of large archives. In some data preservation environments, data loss is just not acceptable for any reason. I often find in these types of environments, when an organization is moving from analog to digital, some managers do not understand that data is not 100 percent reliable on digital media, and having multiple copies of digital media costs more money than keeping books on a shelf, given that data must be migrated to new media, and it is still not 100 percent reliable without many copies of data.

Recommendations for Disk- and Tape-based Archives

I recommend the following data protection policies and procedures for large archives. Except where noted, these recommendations apply to both disk-based archives and tape-based archives.

Data should be synchronously replicated, and validated, to another location that is outside the potential disaster area. For example, if you are in an area that has tornados the replication should be at least 100 miles — or, better yet, 500 miles — north or south of the base location as most tornadoes travel east to west.

Have additional ECC or checksums available to validate the data. Most HSM systems have per-file checksums available on tape, but most do not have them on disk. Technologies such as T10 DIF/PI for tape and disk will become available this year, and many vendors are working on end-to-end data integrity techniques. Per-file checksums are starting to become part of a common discussion in the file system community, but a checksum does not correct the data; it tells you only if the file has gone bad. If you want to know where it went bad within the file you need to have ECC within the file to detect, and hopefully correct, the failure.

In the case of disk-based archives, all RAID devices should have "parity check on read" enabled. Some RAID controllers support this, but others do not. And some RAID arrays support this feature, but it causes significant performance degradation. This feature provides another level of integrity over just having per-file checksums if the failure of the checksum is caused by some failure issue within the storage system. Parity check on read ensures that the failure of a block of data is found on the RAID controller before it is the failure of an entire file.

In the case of tape-based archives, it's important to note that data does not move directly to tape, but to disk and then to tape via HSM. Again, RAID devices should have parity check on read enabled.

Ensure that error monitoring for both soft and hard errors is done on all aspects of the hardware. Soft errors eventually turn into hard errors and, potentially, data failure. Soft errors should be quickly addressed before they become hard errors. This is a significant problem for tape, as there is no standard for Self-Monitoring, Analysis and Reporting Technology (SMART). For more information, see "Solving the Storage Error Management Dilemma."

If possible, *regularly protect and back up the metadata* for the file system and HSM metadata for data on tape because metadata can be restored without restoring all of the data in case of a failure. This works far better, and is far easier, if metadata and data are separated in the file system.

Validate per-file checksums regularly. For large archives, this becomes a significant architectural issue given the CPU, memory and I/O bandwidth required.

DR planning for disk- and tape-based archives is similar. Some of the technologies are different, but the key is regular validation and preparation for the disaster that might come. Far too many organizations do not properly fund large archives and yet expect no data loss. If you have a 50 PB archive and a single replicated site and you lose the archive due to a disaster, you will almost certainly lose data when you re-replicate the site. There is no way to get around the hard error rates in the media. ■

"If possible, regularly protect and back up the metadata for the file system and HSM metadata for data on tape because metadata can be restored without restoring all of the data in case of a failure."

