



INSIGHT

The Copy Data Problem: An Order of Magnitude Analysis

Laura DuBois

Natalya Yezhkova

Ashish Nadkarni

IDC OPINION

Copy data is everywhere. It is on our laptops, it is on our smartphones, and it is most definitely present in large quantities on servers and in the cloud. Creating copies of primary data (i.e., anything that is a copy of the original) is part of human existence — we all want to make copies of anything important. We make copies of data at home and at work, and we even design systems to make copies for us automatically. Protecting data by creating copies provides a sense of security and comfort — for individuals and businesses. We have become so used to creating copies that we hardly worry about how much it actually costs to generate and store these secondary and tertiary copies. That is, until it can be quantified and more importantly demonstrated that there are ways to manage data copies more efficiently.

The copy data problem exists across firms and data centers of all sizes. Copy data exists across backup environments, Disaster Recovery systems, test and development environments, structured and unstructured archiving and big data analytics clusters. For firms, the complex nature and scale of data center infrastructure can make it difficult for IT professionals to quantify how much of a financial and infrastructure overhead copy data really creates. Nevertheless, figuring this out is an important endeavor that serves as an essential first step in making the datacenter infrastructure efficient. Based on the approach presented in this document, IDC estimates that in 2012, more than 60% of enterprise disk storage systems (DSS) capacity may have been made of copy data. Similarly, in 2012, copy data made up nearly 85% of hardware purchases and 65% of storage infrastructure software revenue. These numbers do not include datacenter facilities costs, which can compound this problem even further.

Unless and until businesses acknowledge that this problem is real and needs to be solved once and for all, copy data will become a virus that will plague datacenters for years. In other words, 6 out of every 10 TB of disk storage systems capacity procured and deployed is allocated to copies of primary data. Scaling this copy data capacity spending and infrastructure back can have a profound impact on today's storage budgets. These numbers highlight the need to minimize the number of data copies generated by applications, databases and infrastructure software. As the amount of data and storage infrastructure in a data center continues to grow unabated, businesses must take concrete steps to minimize the sprawl of copy data:

- Aggressively deploying storage optimization technologies like deduplication and compression on all datasets and not just primary data
- Deploying smart copy data management solutions from suppliers like Actifio that automatically manage copy data



- ☒ Becoming smarter about how they approach obvious culprits like what gets put on tape and/or gets replicated

Maintaining copies of data both locally and geographically, and in an online or offline manner is absolutely essential for maintaining service quality. But at some point, it goes from being an insurance policy to paranoia. IDC expects this paranoia to cost businesses roughly \$44 billion in 2013.

IN THIS INSIGHT

This IDC Insight assesses the copy data problem in the datacenter. To assess the problem in a holistic manner, IDC examined the following data sources:

- ☒ Hardware capacity and revenue data from IDC's enterprise disk storage systems 2012–2016 forecast, where IDC looked at data types, usage patterns, and platform types and examined tape revenue and capacity data
- ☒ Software tracker and forecast data for storage infrastructure software and database tools, which were grouped into software that generates copy data, software that is partially responsible for copy data, and software that is influenced by copy data

SITUATION OVERVIEW

Mankind's habit of creating copies of originals is an age-old human trait. It is no surprise that this habit made its way into the digital universe. Humans initially designed systems with a goal of data immortality in mind, and even today leverage some of these same design principles. In other words, no matter what happens, they want their data to be available. Generating copies of data is one of the most common and widely adopted approaches in the digital universe. Examples of how primary data gets proliferated across the datacenter in the form of copies include:

- ☒ **Copies within the system:** To protect against component failure or against human-induced data loss, approaches such as clones and snapshots are often deployed within storage systems. IT administrators often create multiple copies of databases and applications as a means to supplement these in-system copies.
- ☒ **Copies on different systems:** To protect against catastrophic events that can render an entire datacenter offline, approaches such as data replication are often used. Traditionally, data replication has been data agnostic and to keep things simple, many IT administrators replicate entire systems from one datacenter to another. Additionally, local and remote backup to disk and disk-based archiving strategies have been deployed for faster recovery and content based location and retrieval of content.
- ☒ **Offline copies of data:** Replicating data to an offline medium tape has been considered an essential part of any datacenter infrastructure strategy. Because

of how data protection software works in most cases, the same data sets are backed up over and over again — and then sent offsite. In many cases, they include system binaries that not only are the same across multiple servers but also don't change that often. With cloud now augmenting tape, data protection software solutions are getting smarter about only pushing changed data to the cloud but many times lack the intelligence to globally deduplicate data.

- ☒ **Humans generating copies:** Because of the fact that some of the previously mentioned technologies function at a block level, they oftentimes cannot detect unstructured data copies that humans generate. Users may often create multiple copies of files and store them in different folders or directories on a network share. They may send these files to their colleagues or coworkers who may then do the same. Before long, multiple copies of the same file end up on the same network share.

- ☒ **The autonomy of copy data generation:** One of the chief drivers of copy data generation is the autonomy various business units have over their own data sets. Certain business units are willing to dictate the creation of multiple copies under the premise of compliance or service quality, because of the need for comprehensive quality assurance, or simply because they have the budget to do so. Depending on procurement practices, these business units end up paying for such excesses, many times ignorant to the amount of money that they're wasting in the process.

Taking these copy data generation mechanisms into account, IDC sized the magnitude of the copy data market both in terms of spending and capacity.. It leveraged its forecast models and tracker data to estimate what percentage of the total hardware and software revenue could be attributed to copy data sets:

- ☒ **Enterprise disk storage systems capacity and revenue:** IDC breaks down DSS into structured, unstructured, and replicated data and then subsequently adds capacity and revenue for DSS in content depots, public clouds, and non-traditional supplier channels. In 2013, IDC estimates that DSS capacity shipped worldwide will be nearly 100EB. It further estimates that of this capacity, nearly 60% — and nearly 85% of revenue recognized — is for copy data. This means that nearly 61EB of DSS capacity, or roughly \$34 billion in revenue, is directly or indirectly consumed by nonprimary data sets. To arrive at these estimates, IDC used primary research data in which respondents were asked how much copy data they felt existed in their environment.

- ☒ **Storage software:** Storage software is divided into the following functional markets: archiving software, data protection and recovery software, replication software, file systems, storage infrastructure software, infrastructure software, storage and device management software, and other software. Each of these markets influences or directly creates copy data. Accordingly, IDC calculated the revenue impact of each of these markets on copy data. As an example, storage replication software directly creates copy data by virtue of the role it plays in generating copy data when used. To complete the picture, IDC added the impact of revenue from database tools that generate or help generate copies of the database either for backups or for development purposes.

- ☒ **Tape hardware and media:** Other than a small component used for archiving purposes (in which the source data gets deleted after it is archived), almost all tape hardware and media are used for storing copies of data. Businesses that seek to minimize extraneous copy data from the IT environment have to examine the impact of tape in their datacenter. In 2013, IDC estimates that businesses will procure nearly 90EB of tape capacity to make up nearly \$1.3 billion of revenue.

When these components are summed up, the overall picture looks pretty grim. Businesses are spending an extraordinary amount of money on managing and/or maintaining nonprimary data copies. Maintaining copies of data both locally and geographically and in an online or offline manner is absolutely essential for maintaining service quality. But at some point, this maintenance goes from being an insurance policy to paranoia. IDC expects this paranoia to cost businesses roughly \$44 billion in 2013.

FUTURE OUTLOOK

Given that the generators of copy data practically straddle several aspects of datacenter infrastructure, businesses cannot take a pedantic approach to minimizing the number of secondary data copies. From their side, suppliers have to create solutions that tackle this problem in a holistic manner — not solutions that tackle hardware efficiency while ignoring software or vice versa.

While suppliers of storage systems make their own storage solutions become smarter at managing secondary copies, software suppliers are also hard at work maximizing storage space when data copies are generated. Newer versions of software solutions are becoming intelligent about how they handle secondary copies, including unstructured data copies. Examples of such efforts are database backup solutions that don't generate entire copies of data each time.

A new breed of storage start-ups is tackling the copy data independent of the traditional and incumbent storage hardware and software suppliers. These solutions tackle the copy data problem by intercepting the I/O path in an in-band or out-of-band manner, much like storage virtualization solutions. By integrating with applications, they can gain a level of awareness that storage virtualization solutions lack. Such start-ups have the benefit of not being tied to hardware or software technologies that directly manage primary data but rather insert themselves into the I/O path whenever copy data gets generated. This ensures that their solutions do not interrupt or constrain I/O activities related to primary data.

Armed with such technologies, IT organizations should find themselves in a better position to rein in runaway copies of data. They can better educate their constituents on how to control copy data by providing measurable ROI metrics.

Suppliers of copy data management solutions like Actifio should continue to educate the market on the cost-benefit analysis of inserting their solutions in storage environments.

Copyright Notice

This IDC research document was published as part of an IDC continuous intelligence service, providing written research, analyst interactions, telebriefings, and conferences. Visit www.idc.com to learn more about IDC subscription and consulting services. To view a list of IDC offices worldwide, visit www.idc.com/offices. Please contact the IDC Hotline at 800.343.4952, ext. 7988 (or +1.508.988.7988) or sales@idc.com for information on applying the price of this document toward the purchase of an IDC service or for information on additional copies or Web rights.

Copyright 2013 IDC. Reproduction is forbidden unless authorized. All rights reserved.