

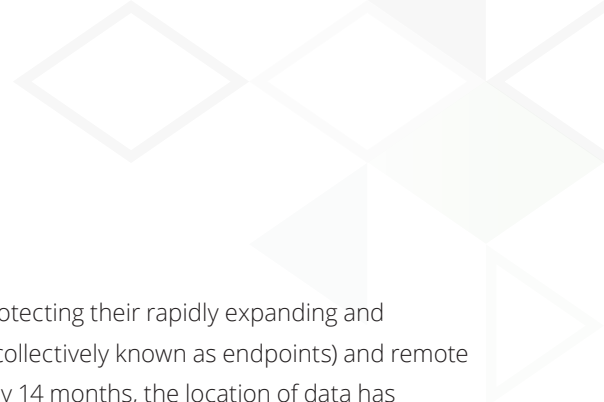
Deduplication and Its Application to Corporate Data



This whitepaper explains deduplication techniques and describes how Druva's industry-leading technology saves storage and bandwidth by up to 80%

Table of Contents

Introduction	3
Increase in Corporate Data is Causing a Sharp Rise in Storage and Bandwidth Costs	3
Data Deduplication Provides Significant Savings in Storage and Bandwidth Costs	3
Different Methods for Data Deduplication	4
Server Side versus Client Side	4
File versus Sub-file Level Data Deduplication	4
Block-based Deduplication.....	4
Application-Aware Data Deduplication.....	5
Understanding Druva’s Industry-leading Deduplication Technology	7
Global Deduplication	8
Client-side Deduplication	8
Object-based Application-Aware Deduplication.....	9
High Performance and Scalable Deduplication	9
Unified Deduplication Across Backup and File Sharing.....	9
Bandwidth and Storage Savings from Druva’s Deduplication	10
About Druva	11



Introduction

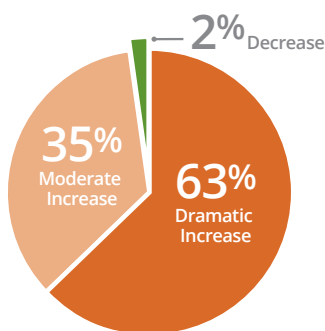
Enterprises are seeking new ways to keep up with the challenges of managing and protecting their rapidly expanding and distributed corporate data—especially data that resides on laptops, mobile devices (collectively known as endpoints) and remote office servers. Growth of data in the enterprise is currently doubling at a pace of every 14 months, the location of data has become more dispersed, and the linkage between data sets more complex. These factors have a significant impact to storage and bandwidth costs. Data deduplication offers enterprises the opportunity to dramatically reduce the amount of storage and bandwidth required for backups and other data storage and access workloads. Druva takes a unique approach to data deduplication technology to help customers address these data challenges.

Increase in Corporate Data is Causing a Sharp Rise in Storage and Bandwidth Costs

Corporate data has increased sharply the past 5 years, which is driving a significant rise in bandwidth and storage costs. A survey by AFCOM (data center trade organization) found that over 63% of IT managers surveyed have seen a dramatic increase in their storage costs.

Two of the main reasons for that increase are the proliferation of mobile devices in the enterprise and the more geographically dispersed nature of enterprises occurring over the last decade. Fortunately and unfortunately for IT managers, much of the data increase is the direct result of replicated files across multiple data repositories and end-point devices.

Change in Storage Requirements Over Past 5 Years



Data Deduplication Provides Significant Savings in Storage and Bandwidth Costs

Data deduplication refers to the elimination of redundant data. Deduplication algorithms identify and delete duplicate, leaving only one copy (or 'single instance') of the data to be stored. However, indexing of all data is still retained should that data ever be needed.

Deduplication is able to reduce the required bandwidth and storage capacity, since only the unique data is stored. For example, a typical email system might contain 100 instances of the same 1 MB file attachment. If the email platform is backed up or archived, all 100 instances are saved, requiring 100 MB storage space. With data deduplication, only one instance of the attachment is actually stored; each subsequent instance is just referenced back to the one saved copy. In this example, a 100 MB storage and bandwidth demand could be reduced to only 1 MB.

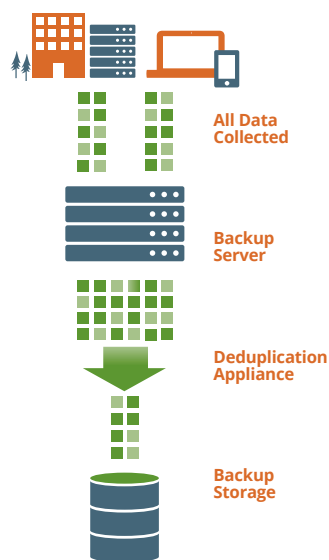
The practical benefits of this technology depend upon various factors, such as point of application, algorithm used, data type and data retention/protection policies. Let's take a look at some of the ways deduplication technologies differ. These technologies differ in the following ways: by where the deduplication happens (server or client side), by granularity of the deduplication (file or sub-file based), and finally by the logic of discovering duplicate data (block-based or app-aware).

Different Methods for Data Deduplication

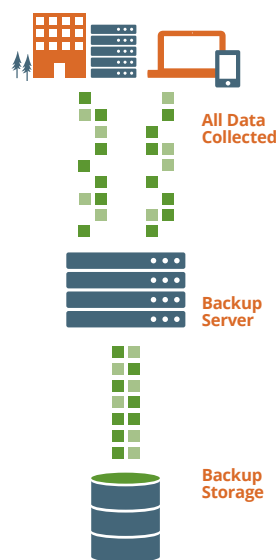
There are several different methods for data deduplication.

Server Side versus Client Side

Server-side Deduplication



Client-side Deduplication



SERVER SIDE BASED DEDUPLICATION method acts on the data on the server. In this case, the client is unaffected and does not benefit from any deduplication.

The deduplication engine can be embedded in the hardware array, which can be used as NAS/SAN device with deduplication capabilities.

Alternatively it can also be offered as an independent software or hardware appliance which acts as intermediary between backup server and storage arrays.

In both cases, this method does not decrease the amount of data transmitted and provides no improvements on bandwidth utilization **it improves only the storage utilization.**

The CLIENT-SIDE DEDUPLICATION METHOD acts on the data at the client i.e., before it is moved to the server. A deduplication-aware backup agent is installed on the client, the agent backs up

only unique data. This approach results in improved **bandwidth and storage utilization.** However, this method imposes additional computational load on the backup client.

File versus Sub-file Level Data Deduplication

The duplicate removal algorithm can be applied on full file or sub-file levels. File-level duplicates can be easily eliminated by calculating a single checksum of the complete file data and comparing it against existing checksums of backed-up files. This approach is simple and fast, but the extent of deduplication is very small, as it does not address the problem of duplicate content found inside different files or data-sets (e.g., emails). The sub-file level deduplication technique breaks the file into smaller fixed or variable size blocks, and then uses standard hash-based algorithms to find similar blocks.

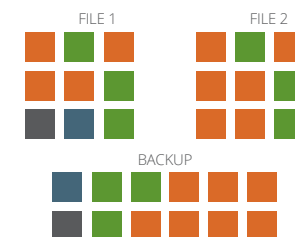
Block-based Deduplication

The block-based deduplication algorithms work the following way: the deduplication engine looks at a sequence of data, segments it into variable length blocks, and seeks blocks that are repeated. The engine stores a pointer to the original block instead of storing the duplicate block again. There are two main types of the block-based approach.

Fixed-length block

approach, as the name suggests, divides the files into fixed-length blocks and uses simple checksum (MD5/SHA etc.) based approach to find duplicates.

Block-based Deduplication



Although it's possible to look for repeated blocks, the approach provides very limited effectiveness since the primary opportunity for data reduction is in finding duplicate blocks in two transmitted datasets that are made up mostly — but not completely — of the same data segments. For example, similar data blocks may be present at different offsets in two different datasets. In other words, the block boundary of similar data may be different.

This is very common when some bytes are inserted in a file, and when the changed file processes again and divides into fixed-length blocks. All blocks appear to have changed. Therefore, two datasets with a small amount of difference are likely to have very few identical fixed-length blocks.

Variable-Length Data Segment technology divides the data stream into variable length data segments using a methodology that can find the same block boundaries in

different locations and contexts. This allows the boundaries to “float” within the data stream so that changes in one part of the dataset have little or no impact on the boundaries in other locations of the dataset.

Through this method, duplicate data segments can be found at different locations inside a file or between different files created by same/different application.

Limitations of Block-Based Deduplication



1 The block size (fixed or floating) used to determine data boundary is usually a “best” guess, & hence may not completely coincide with application’s actual block size.



2 Different applications have different ways of writing on-disk data, block based algorithm will often fail to detect identical blocks across different application file types (e.g. the same block of text stored in MS Word file and in a .PST email file).



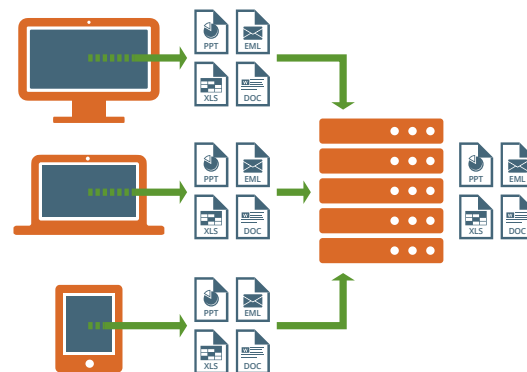
3 Applications like Microsoft Outlook and Office use a complex database based on disk data structure which “stamp” each block with a unique header and footer, further complicating the task of finding duplicate blocks of data.

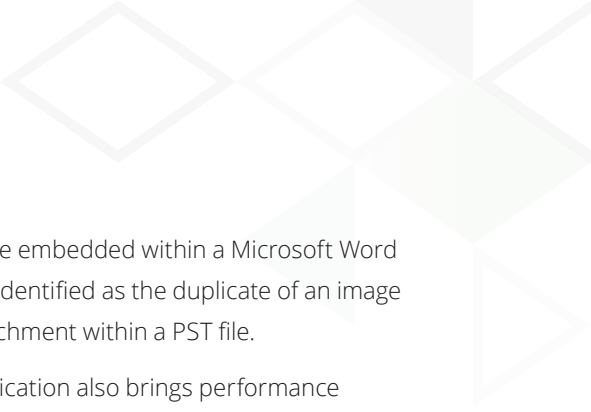
Application-Aware Data Deduplication

Application-aware (App-aware) deduplication is a revolutionary concept which overcomes the limitations imposed by block-level deduplication. It benefits from knowledge of the format of data being backed up. Instead of guessing the optimal block size, it interprets the file as the actual application would, and identifies the logical blocks or messages within the files that have changed.

The deduplication algorithm **removes duplicates at logical block or message level** and is highly accurate as it relies on understanding the structure of on-disk data.

App-aware Deduplication






This approach proves to be highly efficient when it comes to complex applications like Microsoft Outlook and Office, which contribute to over 95% of the data on corporate PCs. App-aware deduplication not only identifies and **removes all duplicates across emails and attachments** within a single PST file, but also effectively **identifies and removes duplicates across different applications**. Using this new


approach, an image embedded within a Microsoft Word document can be identified as the duplicate of an image present as an attachment within a PST file.

App-aware deduplication also brings performance improvements in the speed of deduplication, as there is little or no scanning of data to find the “floating” block boundaries.


Advantages of App-Aware Over Block Based Deduplication



HIGHLY EFFICIENT
in removing duplicates across applications.



UP TO 300% MORE EFFICIENT
than simpler block-based approach in removing duplicates within complex applications like Microsoft Outlook.



UP TO 200% FASTER
data processing compared to the variable block based approach.

EXAMPLE:

Assume a short email message is stored the following way on the disk.

Date 07.01.2014 ; From: Bill Gates ; To: Warren Buffett ; Subject: Sell ; Body: sell everything!

Now assume that when Bill opens Outlook in a week time, and Outlook changes the date to 07/17 storing the message as:

Date 07.17.2014 ; From: Bill Gates ; To: Warren Buffett ; Subject: Sell ; Body: sell everything!

Block based deduplication methods will store the message in the following blocks before and after the change

BEFORE:

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5
Date 07.1.2014 ; Fro	m: Bill Gates ; To: W	arren Buffett ; Sub	ject: Sell ; Body: s	ell everything

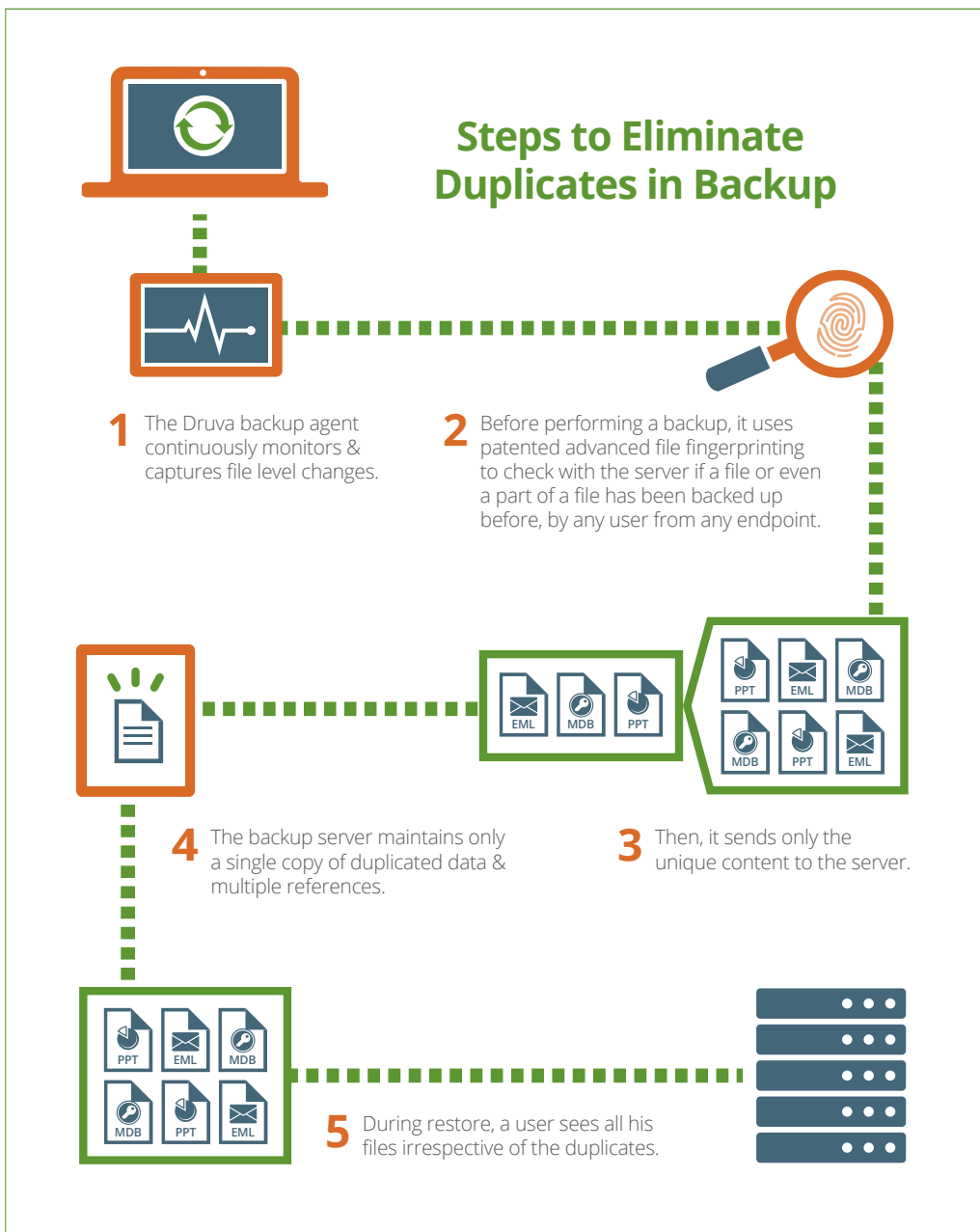
AFTER:

BLOCK 1	BLOCK 2	BLOCK 3	BLOCK 4	BLOCK 5
Date 07.17.2014 ; Fr	om: Bill Gates ; To:	Warren Buffett ; Su	bject: Sell ; Body:	sell everything!

Note that the changing of the day to two characters from one causes all data to shift and all the block to be different. The block based deduplication engine will store a whole new copy of the message even though just one character was changed. Variable block based deduplication does a little better in this case, however, it is still not optimal.

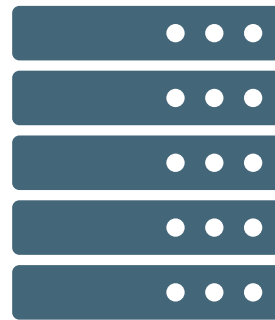
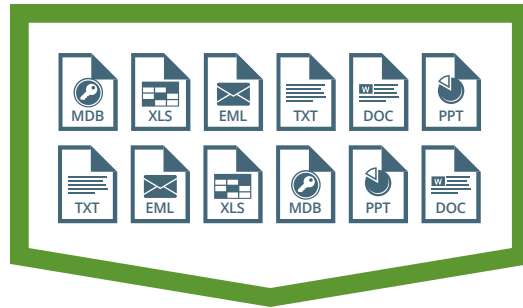
Understanding Druva's Industry-leading Deduplication Technology

Druva's patented global and app-aware deduplication technology provides unmatched bandwidth and storage savings when it comes to both backup and file sharing.



Salient Features of Druva's Deduplication Technology

- 1 Global:** data redundancy is eliminated across all users and endpoints.
- 2 Client-side:** duplication checks and caching of these checks are performed at the client substantially reducing bandwidth and speeding up backups.
- 3 Application Aware:** understanding of on-disk formats of applications, for example in Outlook, results in 100% accurate, faster deduplication and reduced storage requirements.
- 4 High Performance:** Global and app-aware deduplication ensures only unique data is transferred to the server thereby resulting in up to 6x faster initial backups and optimal WAN bandwidth utilization.
- 5 Unified Across Backup & File-Sharing:** deduplication works across backup and file-sharing functions to give massive storage & bandwidth savings



A detailed explanation of deduplication techniques and Druva's innovative approach to deduplication is available in a white paper on data deduplication for endpoints at www.druva.com/resources.

Global Deduplication

The client agent performs duplicate checks at the client device by comparing across data from all enterprise users and from all their devices. Only a single instance of a block from across all devices/users is stored on the server. This combination of Global, and client-side deduplication provides 80% savings in bandwidth and storage.

Client-side Deduplication

Druva uses a client-triggered architecture with deduplication performed at the client enabling high levels of scalability and security. By performing deduplication checks at the client, Druva is able to save substantial bandwidth. Additionally, client-side caching of these file and sub-file level deduplication checks makes backups markedly faster.

The client also has a powerful WAN Optimization Engine, which can automatically prioritize network availability and set backup bandwidth as a percentage of total available bandwidth. By doing so, Druva ensures that a backup neither consumes a large percentage of the bandwidth nor disrupts the end user experience.

Object-based Application-Aware Deduplication

Druva understands the on-disk format of applications and uses this knowledge to significantly enhance the deduplication process while guaranteeing 100% accuracy. The app-aware deduplication technology recognizes common applications such as Outlook (PST). App-aware deduplication eliminates the dependence on multiple checksums resulting in faster deduplication. For other applications, variable-length block based deduplication methodology is used.

DEDUPLICATION WITHIN APPLICATIONS

Many applications tend to change the data structure of a file even if a small element of that data structure changes. As a result, the entire file appears different when stored



persistently on disk. Consider for example, an Outlook PST file. The PST file changes up to 5% of the blocks even if the user closes Outlook without any update. Druva's app-aware technology recognizes up to 87 different message types in PST to intercept the actual changes and back up only unique content (such as emails, attachments, calendar updates, etc.)

This approach guarantees 100% deduplication accuracy on supported applications and optimal use of storage and bandwidth.

DEDUPLICATION ACROSS APPLICATIONS

Each application stores data differently on disk and often the representation of the data changes completely once stored and indexed on disk. A good example is an image file which is present in a Word document as well as in a PST file as an attachment. Block-based deduplication is often unable to identify such duplicates across applications (as the data itself has changed).

Since Druva understands the logical view of the data it is much more efficient in discovering duplicates across applications and eradicating these than other deduplication based backup solutions.

EXAMPLE:

Using app-aware deduplication, a Word document on a user's desktop can be easily identified as a duplicate of an email attachment that has been backed up, and can be removed from the backup.

Unified Deduplication Across Backup and File Sharing

Druva inSync provides both file sharing & collaboration and backup capabilities and the deduplication algorithm works across both of these functionalities further decreasing bandwidth and storage costs.

Organizations typically use two separate solutions for file sharing and backup. As a result, the same file might have multiple copies on the server — stored first by a backup utility and then again by a file-sharing tool. While the file-sharing tool in use might be capable of deduplication, it cannot perform deduplication across files stored by another tool such as a backup utility. Druva inSync is the only solution that allows an organization to use a single tool for both file sharing and backup. As inSync knows which files are stored for backup, and which are for file sharing it can find and eliminate duplications across the two functions resulting in significant bandwidth and storage cost advantages.

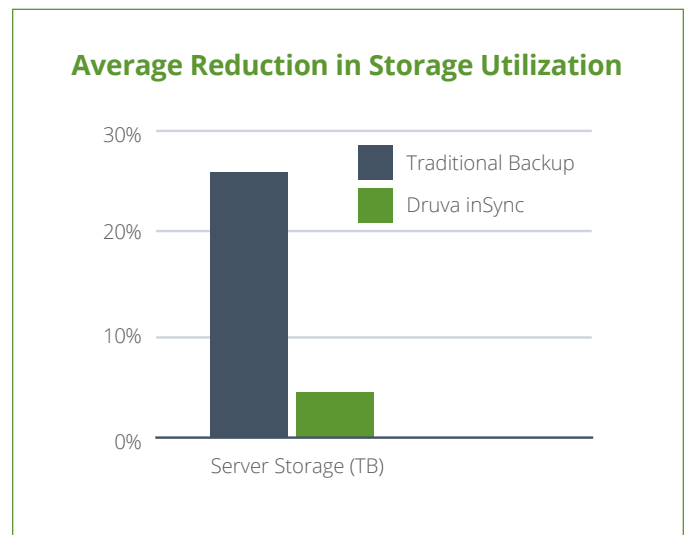
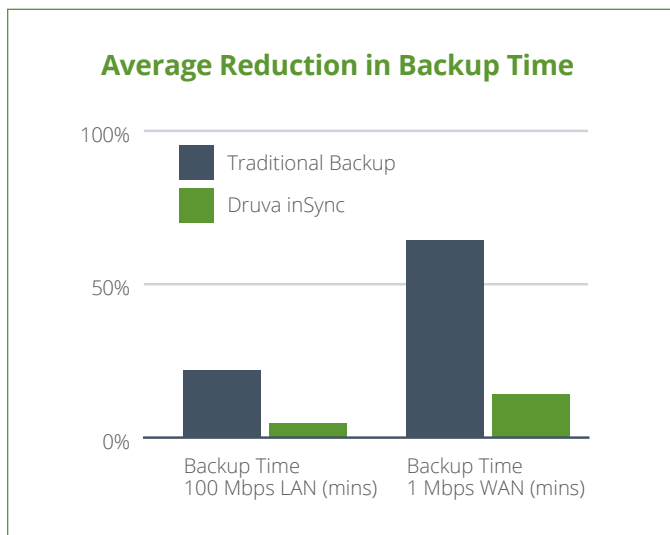
Bandwidth and Storage Savings from Druva's Deduplication

The following table benchmarks the performance of Druva's deduplication technology against incumbent installations at four different customers in different industry verticals. These benchmarks clearly demonstrate the benefits delivered by Druva's global and app-aware deduplication technology in terms of backup time and storage utilization.

EXAMPLE

If an email with 1 MB attachment is sent to 1000 users, traditional incremental backup software would backup this 1 MB attachment from each of the 1000 different mail boxes. Druva inSync would, in contrast, backup 1 MB from the first user, and then skip all the other 999 copies as duplicates, saving over 99.9% backup time, bandwidth and storage.

CUSTOMER	NO OF PCs	AVG. BACKUP TIME LAN (MIN)		AVG. BACKUP TIME VPN/WAN (MIN)		TOTAL STORAGE USED (TB)	
		Old App	inSync	Old App	inSync	Old App	inSync
Large Financial Corp.	2000	24	8	90	20	60	12
Oil & Gas Company	500	10	4	N/A	9	10	1.2
Consult. Group	300	15	6	40	8	27	2
Graphic Company	100	45	14	N/A	6	6.8	1.6



About Druva

Druva is the leader in data protection and governance at the edge, bringing visibility and control to business information in the increasingly mobile and distributed enterprise. Built for public and private clouds, Druva's award-winning inSync and Phoenix solutions prevent data loss and address governance, compliance, and eDiscovery needs on laptops, smart devices and remote servers. As the industry's fastest growing edge data protection provider, Druva is trusted by over 3,000 global organizations on over 3 million devices. Learn more at www.druva.com and join the conversation at twitter.com/druvainc.



Druva, Inc.
Americas: +1 888-248-4976
Europe: +44(0)20.3750.9440
APJ: +919886120215
sales@druva.com
www.druva.com