Whitepaper – Data Deduplication and Druvaa inSync

Whitepaper

The whitepaper explains source based data de-duplication technology and how it is used by Druvaa inSync product to save storage and bandwidth.

Druvaa Software

docs@druvaa.com

WP/100 / 009

12/31/2008

Table of Contents

Introduction

Enterprises are seeking new ways to tackle their data protection challenges. While data growth is not new, the pace of growth has become more rapid, the location of data more dispersed, and the linkage between data sets more complex. Data deduplication offers companies the opportunity to dramatically reduce the amount of storage and bandwidth required for backups. Druvaa inSync uses data deduplication technology to help customers address these data protection challenges.

With over 80% corporate data duplicate across PC users, this whitepaper discusses how source based data deduplication can save up to 90% bandwidth and storage when compared to traditional backup methods.

Towards the end, the whitepaper discusses in details how Druvaa inSync with its client triggered backup architecture and unique data de-duplication approach, makes backup almost invisible for local and remote users.

Understanding Data Deduplication

Data deduplication or Single Instancing essentially refers to the elimination of redundant data. In the deduplication process, duplicate data is deleted, leaving only one copy (single instance) of the data to be stored. However, indexing of all data is still retained should that data ever be required.

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 megabyte (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.*

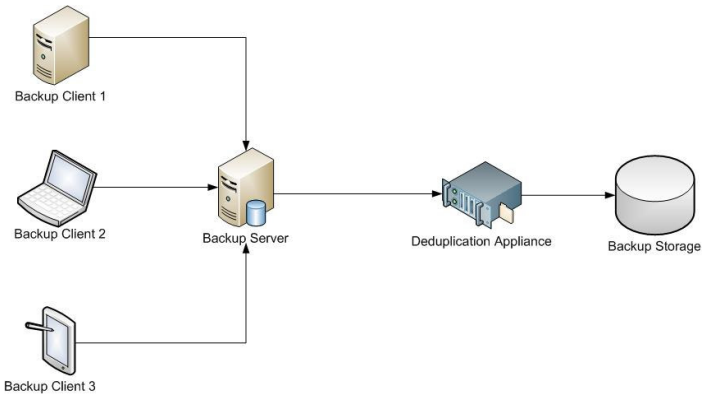
The practical benefits of this technology depend upon various factors like – point of application, algorithm used, data type and data retention/protection policies. Let's take a look at some of the key technology differentiators.

Target vs Source based Data Deduplication

Target based deduplication acts on the target data storage media. In this case the client is unmodified and not aware of any deduplication.

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

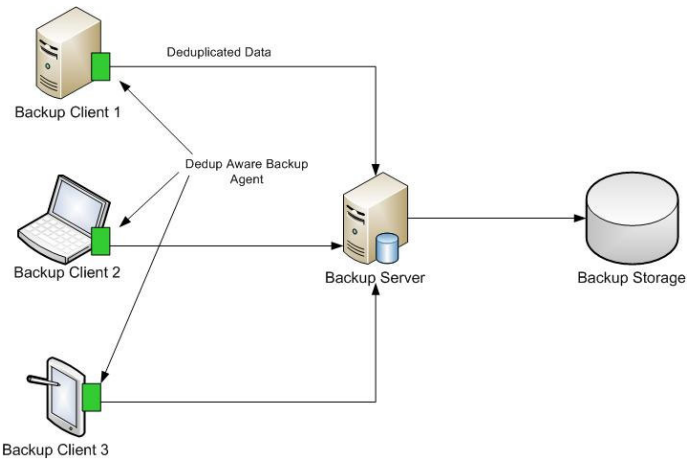
Alternatively it can also be offered as an independent software or hardware appliance which acts as intermediary between backup server and storage arrays.



In both cases it improves only the storage utilization.

On the contrary **Source based deduplication** acts on the data at the source before it's moved. A deduplication aware backup agent is installed on the client which backs up only unique data.

The result is improved *bandwidth and* storage utilization. But, this imposes additional computational load on the backup client.



File vs Sub-file Level Data Deduplication

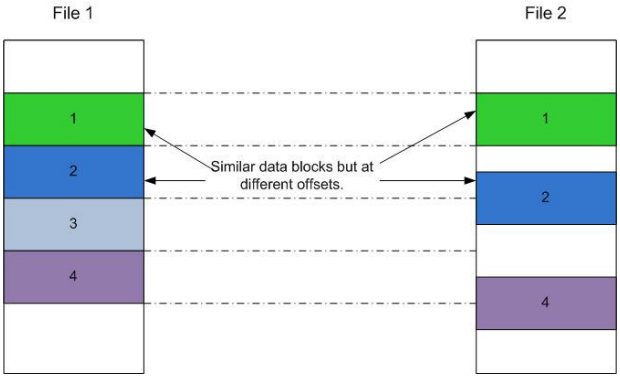
The duplicate removal algorithm can be applied on full file or sub-file levels. Full file level duplicates can be easily eliminated by calculating single checksum of the complete file data and comparing it against existing checksums of already backed up files. It's simple and fast, but the extent of deduplication is very less, 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 algorithm to find similar blocks.

Fixed-Length Blocks v/s Variable-Length Data Segments

Fixed-length block approach, as the name suggests, divides the files into fixed size length blocks and uses simple checksum (MD5/SHA etc.) based approach to find duplicates. Although it's possible to look for repeated blocks, the approach provides very limited effectiveness. The reason is that 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, inside different files, inside files created by different applications, and inside files created at different times.

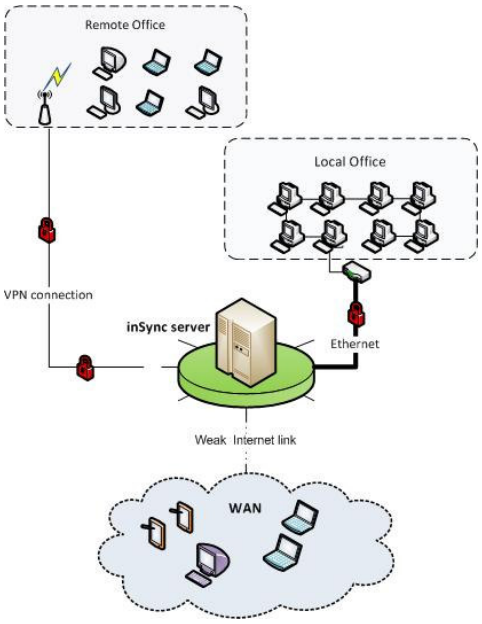
Understanding Druvaa inSync and Source-based Deduplication

Druvaa inSync is an automated enterprise laptop backup solution which protects corporate data while in office or on-the-move. It features simple backup, point-in-time restore, and patent-pending de-duplication technology to make backups much faster.

Architecture – Client Triggered Secure Backups

Druvaa inSync architecture has two components –

1. inSync client and
2. inSync enterprise server.



Druvaa inSync client is a host based soft driver which gets installed on the user PCs. It is equipped with sufficient backup intelligence to *initiate and accomplish backup*. Configuring a client is a simple 5 step effort and can be completed within minutes of installation. The client triggered backup architecture enables high levels of scalability and security.

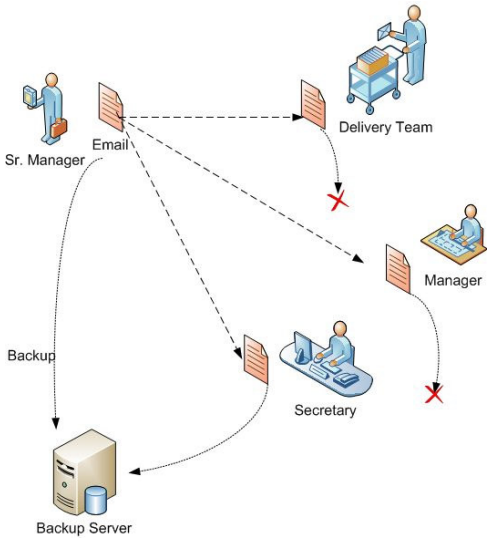
The client also has a powerful WAN optimizer to automatically prioritize network and schedule backup bandwidth as a *percentage* of available bandwidth.

Druvaa inSync Enterprise Server is a software service which runs on a dedicated sever and can scale to serve terabytes of enterprise data. The server accepts backup and restore requests on published IP addresses using a *256-byte SSL encrypted* channel and stores it locally on a *256-bit AES encrypted* storage.

The server offers intelligent user and storage capacity management which enable the administrator to create and control centralized backup policies. Advanced reporting offers various alerts, server health and user statistics. This makes the task of managing remote users much simpler.

Source based Data Deduplication

Druvaa inSync uses patent pending **data de-duplication** technology to remove duplicate data at the source (user's PC) before the actual data backup is initiated. The deduplication algorithm uses **variable-length data segment** approach to find sub-file level duplicates across users.



For example, an email follow up within a team creates 10 copies of same attachment. The deduplication aware backup agent, checks for duplicates with server. The server maintains single copy with multiple references.

With each backup data recorded and indexed, this enables **time-line based *from-the-past restore***. In case of loss of data the user can start a restore from multiple point-in-time restore points.

Bandwidth and Storage Savings

Each organization has a capacity to generate data. The extent of savings depends upon – but not directly proportional to – the number of users. Overall the deduplication savings depend upon following parameters –

1. No. of users
2. Avg. Data / PC and avg. daily change
3. Type of data (emails/ documents/ media etc.)
4. Backup policy (weekly-full – daily-incremental or daily-full)
5. Retention period

Based on these parameters, following are benchmarks obtained from some customers –

Customers	PC Users	Avg. Data /PC (GB)	Pec. Change	Data Type	Backup Policy	Retention Period	Inc. Backup Time over LAN (mins)		Total Storage (TB)	
							Old App	inSync	Old App	inSync
Large Financial Corp.	1000	20	5	Docs & Emails	Daily-inc Weekly-full	90 days	24	3	350	18
Oil and Gas company	500	6	2	Mostly Emails	Daily-inc Weekly-full	30 days	8	1	14	1.2
Consultancy Group	300	10	2	Docs & Emails	Daily-inc Weekly-full	90 days	5	1	43	2
Small Graphic Design Company	100	45	1	Mostly Media	Daily-inc Weekly-full	15 days	24	3	6	2

Customer statistics have helped us to conclude that the following savings are easily achievable when compared with traditional backup methods –

1. Storage savings - **1:21**
2. Backup time savings –
   a. Over LAN - **1:10**
   b. Over WAN – **1:30**
3. Overall TCO **– 1:20**

Druvaa inSync – Ideal for Enterprise PC Backups

Druvaa inSync is designed keeping "mobility" in mind. To summarize, some of the key product highlights which make inSync ideal for on-the-move or remote backups -

1. **10X Performance -** Druvaa inSync uses the patent pending at-source data de-duplication technology to cut down duplicate data and deliver up to 10X faster backup speeds at 90% reduction in bandwidth and storage.

2. **Client trigged Backups -** Client triggered backups ensure that backup/restore requests are initiated by the client and no requests goes out from the server. This has the following benefits –
    a. Users can easily do a backup over WAN
    b. Backup server is secure
    c. Improved scalability

3. **Secure -** Druvaa inSync uses 256-byte SSL encryption for network communication and 256-bit AES encryption for storage.

4. **WAN Optimization -** The inSync client automatically senses changes to the network and makes changes to backup bandwidth and packet size. On-wire compression further ensures optimal utilization of bandwidth.

5. **Browser Based Restores -** When not on one's PC, the user can use browser to access the backed up data over HTTPS.

6. **Advanced Reporting -** The administrator can remotely monitor user activity and help him troubleshoot. The inSync server also offers six different reports for extensive reporting.

Some of the other important features –

1. **Usability**
    a. **Easy**, automated installation and transparent non-intrusive backups.
    b. **Opportunistic Scheduling** starts sync on availability of bandwidth.
    c. **Intuitive graphical interface** to manage and monitor backup.
    d. **Locked/Open File Support** for files like Outlook working files (PST files).

2. **Administration**
    1. **User Profiles** facilitates the administrator to view/guide/control users configuration
    2. **Manage storage capacity** and user quota
    3. **Live server health** and user backup statistics.
    4. **Configurable trigger based reporting** enables queries for relevant information
    5. **Email notifications** for detailed reports

About Druvaa

Druvaa provides enterprise class consultancy and solutions for data availability and business continuity. Information about Druvaa can be obtained from www.druvaa.com