Dell EMC Unity: Data Reduction Analysis

Data reduction on application-specific datasets

Abstract

This document analyzes Dell EMC™ Unity data reduction ratios for various application-specific data types to help administrators accurately estimate space savings.

February 2019

Revisions

Date	Description
January 2019	Initial release
February 2019	Updates to section 2.3

Acknowledgements

Author: Darin Schmitz

Table of contents

Executive summary

Data reduction techniques such as compression and deduplication within storage arrays have been indispensable to help reduce storage consumption. Unfortunately, due to the variability of data types stored within any given storage array, data reduction savings can greatly fluctuate between environments, making it difficult for administrators to estimate storage needs.

This document explores potential savings using real-world data reduction ratios on a Dell EMC™ Unity array. Data was gathered from common applications such as VMware® virtual machines, file-share data, Microsoft® SQL Server® databases, Microsoft Hyper-V® virtual machines, and more.

1 Introduction

When exploring compression and deduplication savings in a storage array, it is important to understand that the types of data stored have a large impact on how reducible the data is. For example, while a volume full of ASCII text documents may have a high reduction ratio, another volume full of MP4 video files may produce very little savings since the audio and video codecs have already reduced the file sizes.

Note: The results presented in this document should only be used as a general guideline. Every environment is different and results may vary based on a number of factors including dataset, workload, or environment. Review the resources outlined in section 1.3 for details and best practice guidelines. For assistance with sizing or designing a system, see your Dell EMC representative.

1.1 Environment and setup

The environment used in this document is composed of a Dell EMC Unity 550F array with 15 x 7 TB flash drives. This array is connected by Fibre Channel to a separate non-Unity array. The configuration simulates migration of data from a different block-based platform to a Dell EMC Unity array to examine potential real-world savings.

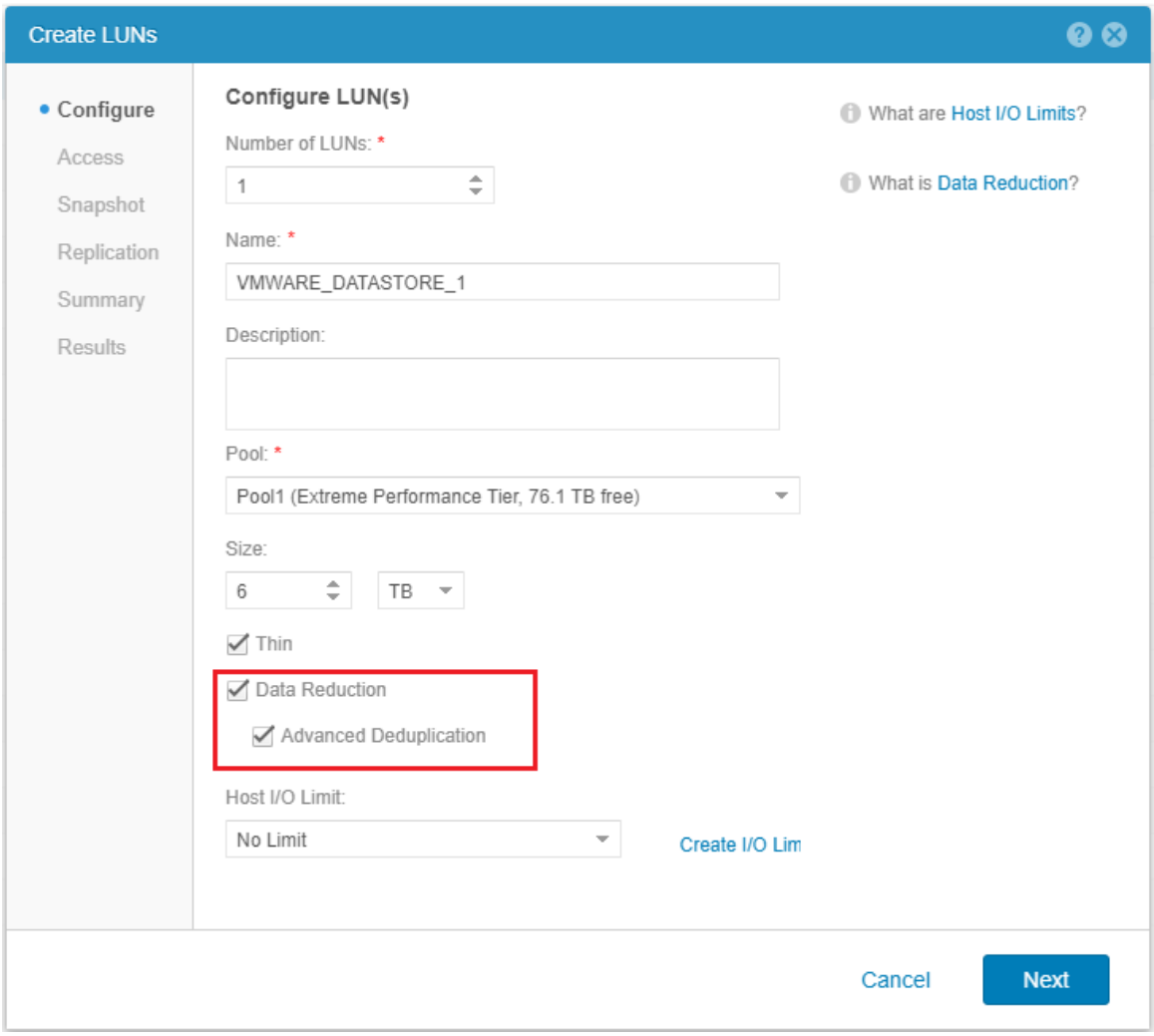

Figure 1 Enabling Advanced Deduplication on the LUNs in Dell EMC Unisphere

1.2 About the test data

To best gauge data reduction savings, this analysis uses real-world data instead of artificially generated data from prevailing test tools. The data is copied to the Dell EMC Unity array from another block-based array. This array stores data for a mid-sized department of technical employees in an existing Fortune 500 company. Due to the size of the department, the data represents a large variety of data types, with several years' worth of active and historical data accumulated.

1.3 Performance reporting

This analysis measures only the data reduction ratios for the various data types. Array performance is out of scope for this document and is not detailed.

For more information regarding data reduction, refer to the following documents:

- Dell EMC Unity: Data Reduction
- Dell EMC Unity: Best Practices Guide

2 Analyzed data types

This analysis takes data reduction measurements from numerous volumes totaling over 35 TB of allocated space and consuming several terabytes of data. The volumes are grouped with similar data types and are separated into the six categories outlined in the following subsections.

2.1 VMware vSphere virtual machines

The virtual machine (VM) datastores tested consisted of various operating systems and applications. The storage-reduction ranges for all VMware VM volumes were as follows:

- Small savings = 1.66:1 (39.89%)
- Average savings = 2.13:1 (51.00%)
- Large savings = 2.97:1 (66.33%)

Small savings: The datastore storing VM templates and ISOs produced the smallest reduction ratio. This result is likely caused by the datastore containing a very small number of VM templates stored alongside a high number of operating system ISO images. It is worth nothing that while the ISO format itself is not compressed, it is common for data contained within ISO images to already be compacted.

Average savings: The average savings figure represents the average reduction ratio and savings percentage across all ten of the VMware datastores analyzed.

Large savings: The largest savings observed was on the datastore where Microsoft Windows Server® VMs were cloned five times from a single template and then powered on. This favorable reduction ratio was achieved because all VMs were cloned from the same template.

Another notable observation regards VMDK formats. To analyze the difference between VMDK formats, one datastore contained five virtual machines in the thick format, and another datastore contained five virtual machines in the eager zeroed thick (EZT) format. Interestingly, this made an almost imperceptible difference in the reduction ratios. The thick-formatted VMs only had a reduction savings of 300 MB more than the EZT-formatted virtual machines (46.2 GB compared to 46.5 GB).

2.2 Microsoft Hyper-V virtual machines

For this analysis, the Hyper-V VM formats were stored in two Cluster Shared Volumes (CSV), each containing a different format of virtual hard disks (VHDs). The first volume stored five VMs with fixed VHDs, and the second stored five VMs with dynamically expanding VHDs. The storage-reduction ranges were as follows:

- Fixed-format VHDs = 1.84:1 (45.78%)
- Dynamic-format VHDs = 1.85:1 (46.08%)

Much like the previous observations of the VMware virtual disk formats, the Hyper-V disk formats also produced a negligible reduction in savings.

2.3 File share

There were three volumes analyzed for file-share data. The first volume analyzed was a 6 TB Windows Storage Server 2016 volume formatted with NTFS. This volume housed multiple SMB file shares, and contained unstructured departmental data such as user directories, a public directory, departmental projects, and a video share.

The second volume was an exact file-copy replica of the first volume but with one difference: It had the Windows Server 2016 Data Deduplication role enabled and set to the **General purpose file server** setting. This tested the effect of software-based deduplication on the array-based data-reduction ratios.

The third volume contained a dedicated application share that stored transactional log files that were output from a specialized Java® application.

The storage-reduction ranges were as follows:

- Small savings = 1.06:1 (5.69%) (Windows Data Deduplication role enabled)
- Average savings = 1.29:1 (22.76%)
- Large savings = 3.14:1 (68.19%)

The average savings ratio is what most administrators may expect to see with traditional Microsoft Windows user shares. Within each of the directories, there were a wide variety of Microsoft Office files, downloaded zip files, executable files, JPG images, Adobe PDFs, and other file types. It is important to remember that many modern file types store data in a reduced format, and many files saved from the internet are also likely to be pre-reduced to save web-server bandwidth.

The small savings ratio was achieved on the volume stored with the Windows Data Deduplication role enabled. Since this data was deduplicated first by the operating system, the array-based deduplication produced very little additional savings.

With the large savings ratio, the share stored an archive of proprietary application log files. These log files contained a large sum of highly reducible data due to the structured format of the output. As with many transactional files, data was written sequentially with numerous timestamped filenames.

Note: It is not a best practice to use array-based deduplication in addition to software deduplication. The results presented in this section are made for illustrative purposes only.

2.4 ISO share

This NTFS volume contained a directory purely consisting of ISOs to analyze the data reduction ratios when measured by themselves. As noted previously, ISOs usually contain compacted data and the low reduction levels were consistent with expectations.

- ISO share = 1.1:1 (9.17%)

2.5 Virtual desktops

This volume contained VMware virtual desktops running Windows 10 and stored them as full clones. Virtual desktops stored as linked clones were not analyzed because deduplication would have been non-productive.

- Virtual desktops = 1.52:1 (34.29%)

This data reduced well because the full-clone VMs contain a lot of duplicate data such as the operating system, Microsoft Office applications, and other productivity tools. However, it is worth noting that within virtual desktops there can still be a lot of unique data in the user profiles.

2.6 Microsoft SQL Server

The Microsoft SQL Server data was divided into two volumes based on Microsoft best practices. One volume was used to store the database files, and the other volume was used to store the transaction logs.

- Database volume = 1.49:1 (32.96%)
- Logs volume = 12.9:1 (92.25%)

Reducing database volumes can produce storage savings, though this falls outside the performance considerations of whether database volumes should have deduplication enabled on them or not. While actual database reduction levels can vary wildly based on the data being stored, the transaction logs were reduced significantly.

3 Conclusion

This document demonstrated data reduction with several different data types, and almost all of them had beneficial savings. Although these savings can be difficult to predict due to the natural variability of data, these test results can help administrators make educated guesses regarding data reduction. However, until data is actually stored on the Dell EMC Unity array, the actual data reductions will be unknown.

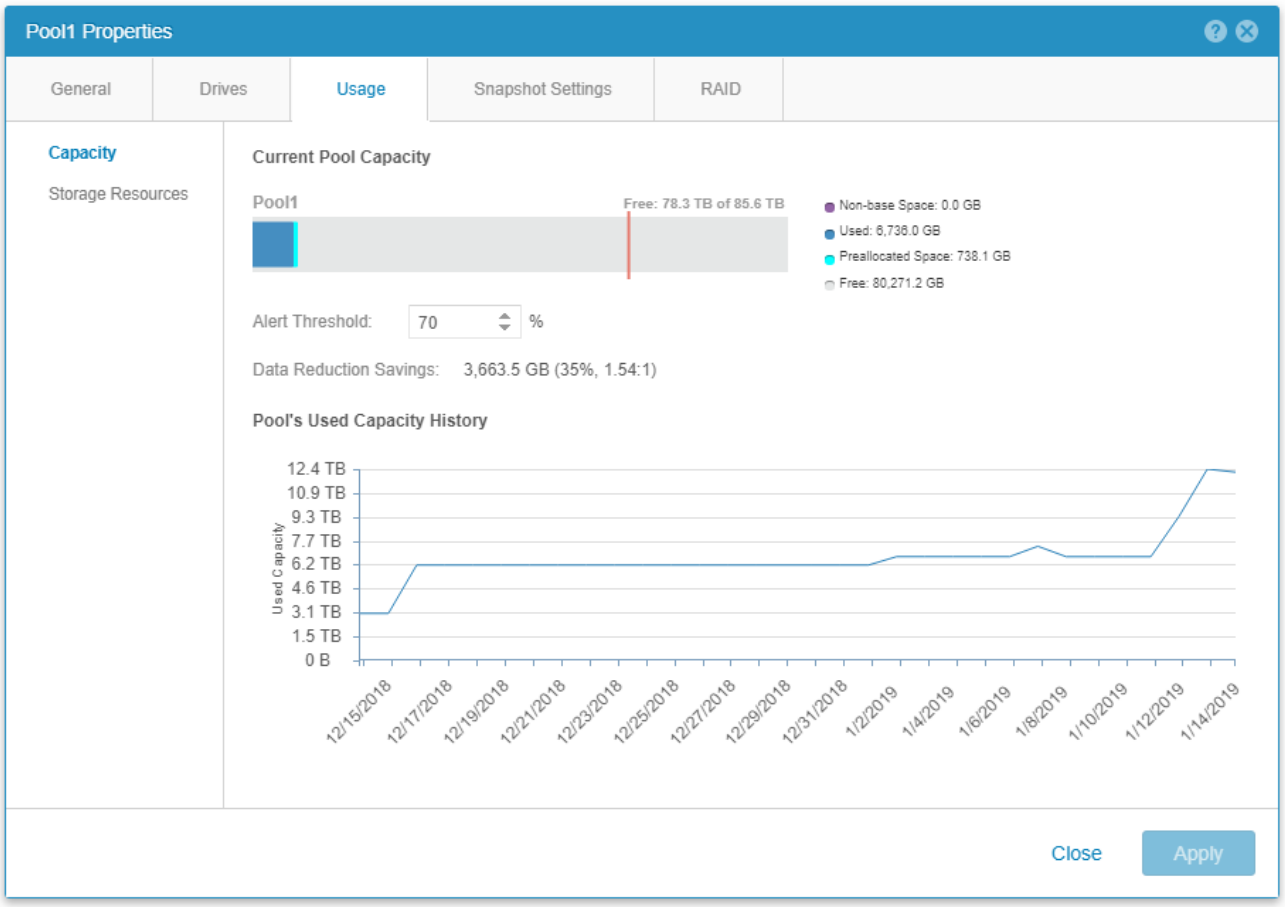

Figure 2 Overall pool savings from the array where all volumes have Advanced Deduplication enabled

In this analysis, the pool-level savings had a definite impact on the bottom-line consumption, but it is important to remember that data reduction is not appropriate for all volumes. Since it is highly unlikely that data reduction will be enabled on every volume in an array, the pool-level results presented are for illustrative purposes only.

A Technical support and resources

Dell EMC Support is focused on meeting customer needs with proven services and support.

Storage technical documents and videos on Dell.com provide expertise that helps to ensure customer success on Dell EMC storage platforms.

A.1 Related resources

Refer to the following referenced or related resources:

- Dell EMC Unity: Introduction to the Platform
- Dell EMC Unity: Best Practices Guide
- Dell EMC Unity: Data Reduction