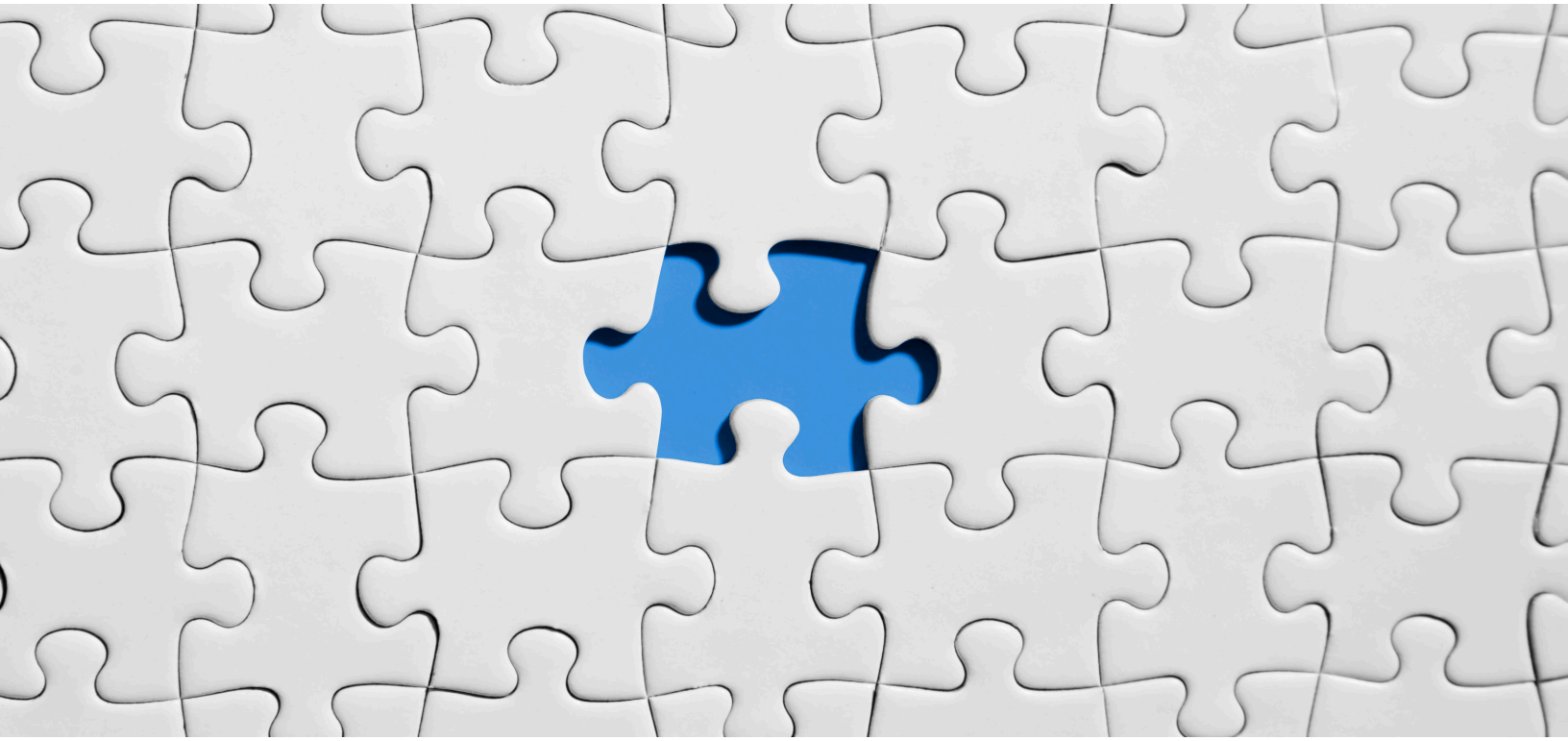
DELL Technologies

DATA DOMAIN SIZING BEST PRACTICES



Pragathi R S

Specialist 2, Inside Product

Dell Technologies

Pragathi.r_somashekar@emc.com

Shilpa Bhasker

Sales Engineer Analyst

Dell Technologies

Shilpa.bhasker@dell.com

Mohammed Moueen Wahab

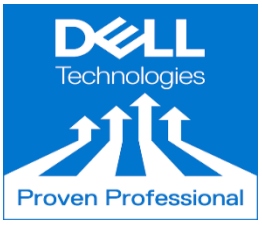
Specialist 2, Inside Product

Dell Technologies

Mohammed.moueen@dell.com

DELL Technologies Proven Professional

Knowledge Sharing Article



The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Introduction

Data growth in backups is a reality as business needs change. In any backup environment it is important to size a solution with accurate capacity your customer would require to meet backup storage capacity requirements. Insufficient capacity and bandwidth to perform the backup can cause backups to lag, fail to complete or fill a backup device sooner than expected. Data Domain is a disk-based inline deduplication purpose built backup appliance that provides data protection and disaster recovery (DR). There are multiple factors to consider such as data type, data size, data reduction rate(deduplication), retention, backup schedule, etc. which affect the usable capacity of Data Domain.

This article explores how these factors affect the capacity requirements of Data Domain, how to use the Solution Builder sizing tool to calculate the usable capacity of the Data Domain device, how to avoid sizing for incorrect usable capacity and the importance of Blue-Green charts for sizing. The main goal is to design the protection storage system with the accurate model and configuration that can store the required front-end data for the given customer requirements.

Audience

This article is intended for Dell Technologies sales and presales field personnel interested in gaining a high-level understanding of Data Domain and basic methodology /process /procedures to size a Data Domain appliance to best meet customer requirements and understand the core metrics that impact its performance.

Data Domain Overview

Data Domain is disk-based Backup and Restore appliance, that provide data protection and disaster recovery (DR) in the enterprise environment. Data Domain is designed to reduce the amount of disk storage that is required to retain and protect data. All systems run on the Data Domain Operating System (DD OS), Data Domain system features ensure data integrity, reliable restoration, efficient resource usage, and ease of management.

A main strength of Data Domain systems is that databases, Email servers, VM's, Enterprise applications and other data sources and a broad range of backup and archival use cases can be protected on a single system. One of the key differentiators Data Domain systems offer is the ability to deduplicate and encrypt data inline as it is written to disk, providing a fast and secure solution.

On the archival side, Data Domain systems can meet a variety of US and international compliance regulations for archive data with retention lock technology. Retention Lock protects selected files from being modified, over-written or deleted before a specified retention period expires.

All data sent to a Data Domain system can be efficiently replicated to a secondary site for disaster recovery or to public, private or hybrid cloud for long-term retention.

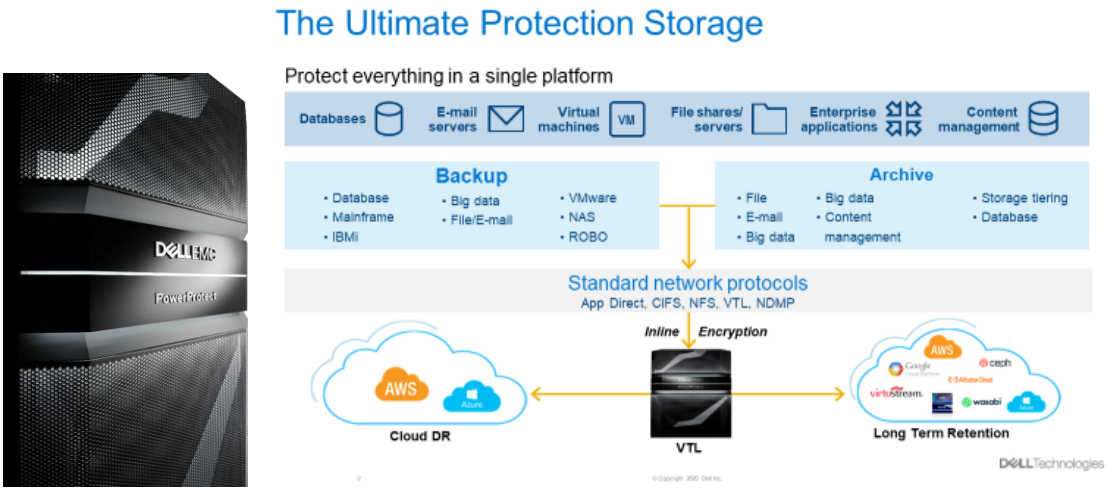


Figure 1 - Data Domain

Inputs for Sizing

What is Sizing?

Creating or developing a solution based on customer requirements.

Sizing is performed based on analyzing, observing and gathering customer data from one or more sources of information.

Sizing Terminologies

- **Site:** Location of the data center or place where the solution would be installed.
- **Front-End Data:** The size of the primary data being protected currently by client applications, virtual machines, and systems. This would be the size of data you would want the solution to protect.
- **Period of Interest:** The amount of time for which the solution is designed. Typically, 3- 5 years.
- **Data Type:** Refers to the type of workload/source where the data is stored and what kind of data it is. Example: VM, DB, Filesystem, NAS.
- **Backup Schedule**: The number of times you want to perform a backup. It could be Full or incremental backups in a week.
- **Retention:** How long do you want to retain backup data on the device/storage. Example: number of weeks/days/ years or any combination of them.
- **Growth Rate (%):** The percentage change of a specific variable within a specific time period. Example: The amount of data that has changed annually. It can be 5%, 10%, 20%, etc.
- **Daily Change Rate (%):** The rate at which the data received from the backup application changes from backup to another backup, i.e. the amount of data that has changed from day 1 to day 2).
- **Backup Window:** The time slot/window most suitable to back up data. It is a predefined/prescheduled time when the backup software is permitted to start the backup process on a computer system.
- **Initial Full Reduction:** Amount of data that is reduced during the initial first full backup.
- **Sub Full Reduction:** Amount of data that is reduced subsequently after the initial first full backup. (depends on backup schedule: Daily full backups)
- **Incremental Reduction**: Amount of data that is reduced incrementally after the initial first full backup. (depends on backup schedule: Weekly full and daily incremental backups)

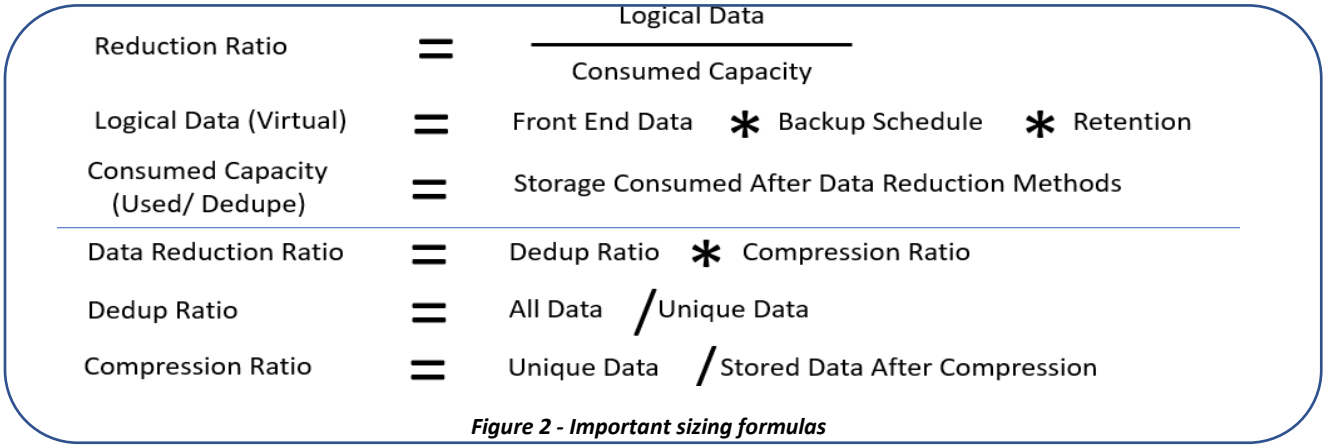


Figure 2 - Important sizing formulas

4.3 DD ASUP and Fifer Script

1. **DD Auto Support (ASUP)**: Auto generated text file created by Data Domain OS and sent to Dell Technologies via e-mail or ESRS at a scheduled time each day.

```
Message-Id: <5c0139$4j4enm>
Date: Tue, 19 Jan 2021 07:12:33 +0000
To: autosupport
User-Agent: Data Domain SMTP Interface
Subject: scheduled autosupport (btcbosdd02)
MIME-Version: 1.0
Content-Type: text/plain; charset="US-ASCII"
Content-transfer-encoding: 7bit
X-autosupport: true
X-autosupport-hostname: btcbosdd02
X-autosupport-systemid: FLA00150400006
X-autosupport-location:
X-autosupport-time: 1611040354
X-autosupport-type: autosupport
Origin: external


==========  GENERAL INFO  ==========
GENERATED_ON=Tue Jan 19 06:56:05 UTC 2021
GENERATED_EPOCH_TIME=1611039365
TIME_ZONE=us/eastern
VERSION=Data Domain OS 6.0.2.20-587212
SYSTEM_SERIALNO=FLA00150400006
CHASSIS_SERIALNO=FLA00150400006
MODEL_NO=DD2500
HW_REVISION=1
DATA_ENCRYPTION_SUPPORTED=Yes
SSD_SHELF_PRESENT=NO
HOSTNAME=btcbosdd02
LOCATION=
ADMIN_EMAIL=
HA_ENABLED=false
UPTIME= 06:56:18 up 823 days, 12:57,  0 users,  load average: 7.11, 6.15, 5.93
Filesystem has been up 823 days, 11:58.

Proactive Disk Check
```

Figure 3 - Sample DD ASUP report

2. **Fifer Script:** These scripts save time and provide more in-depth data than Validate the Value.

Metric	Peak Value	Date
Active Tier Total Post-Comp (TB)	17.44	04/07/2020 03:07 AM
Active Tier Total Pre-Comp (TB)	159.25	04/06/2020 03:01 AM
Active Tier Total Reduction (X:1)	10.46	04/01/2020 03:00 AM
Active Tier Utilization	29%	04/07/2020 03:07 AM
Active Tier Daily Pre-Comp (TB)	19.14	03/22/2020 05:43 AM
Active Tier Daily Post-Comp (TB)	1.46	04/05/2020 02:59 AM
Active Tier Daily Reduction (X:1)	27.60	04/07/2020 03:07 AM
Cloud Tier Total Post-Comp (TB)	N/A	N/A
Cloud Tier Total Pre-Comp (TB)	N/A	N/A
Cloud Tier Total Reduction (X:1)	N/A	N/A
Cloud Tier Utilization	N/A	N/A
Cloud Tier Daily Pre-Comp (TB)	N/A	N/A
Cloud Tier Daily Post-Comp (TB)	N/A	N/A
Cloud Tier Daily Reduction (X:1)	N/A	N/A
Read Streams	4	03/29/2020 08:50 PM
Read+ Streams	N/A	N/A
Write Streams	41	03/22/2020 01:00 AM
Write+ Streams	N/A	N/A
Replication In Streams	N/A	N/A
Replication Out Streams	N/A	N/A
Total Streams	42	03/22/2020 01:00 AM
Read Throughput (MB/s)	N/A	N/A
Write Throughput (MB/s)	1,102.20	04/05/2020 01:10 AM
Replication Pre-Comp In (MB/s)	N/A	N/A
Replication Pre-Comp Out (MB/s)	N/A	N/A
Replication Network In (MB/s)	N/A	N/A
Replication Network Out (MB/s)	N/A	N/A
Active Mtrees - Read	1	04/05/2020 03:50 PM

Index | KPIs | General | Filesystem | Concerns | Current Alerts | Feature Licenses | Capacity Licen ...

Figure 4 - Sample Data Domain Fifer Script

Data Domain Sizing Tool

Solution Builder Overview

Solution Builder, a Data Domain sizing tool, is a web-based application accessible anywhere on any device and is widely used by Data Protection presales/sales team to design solution for their customers. Solution Builder's interface will help you enter customer data and enable you to review potential solutions using different combinations of Data Protection Specialty (DPS) and Dell Technologies products and compare configurations. It is designed to reuse common information about customers' environments across all the products contained in the solution. By doing this, the user only needs to enter the information that is unique to the product being sized.

Solution Builder Tool Compatibility

Solution Builder supports sizing of the Data Domain, IDPA, PowerProtect with Data Domain, Avamar with Data Domain, NetWorker with Data Domain, Data Domain with third-party backup and recovery application, Avamar Data Store, Data Protection Advisor, Protect Point for VMAX and CloudBoost Connector.

Exploring Solution Builder Options

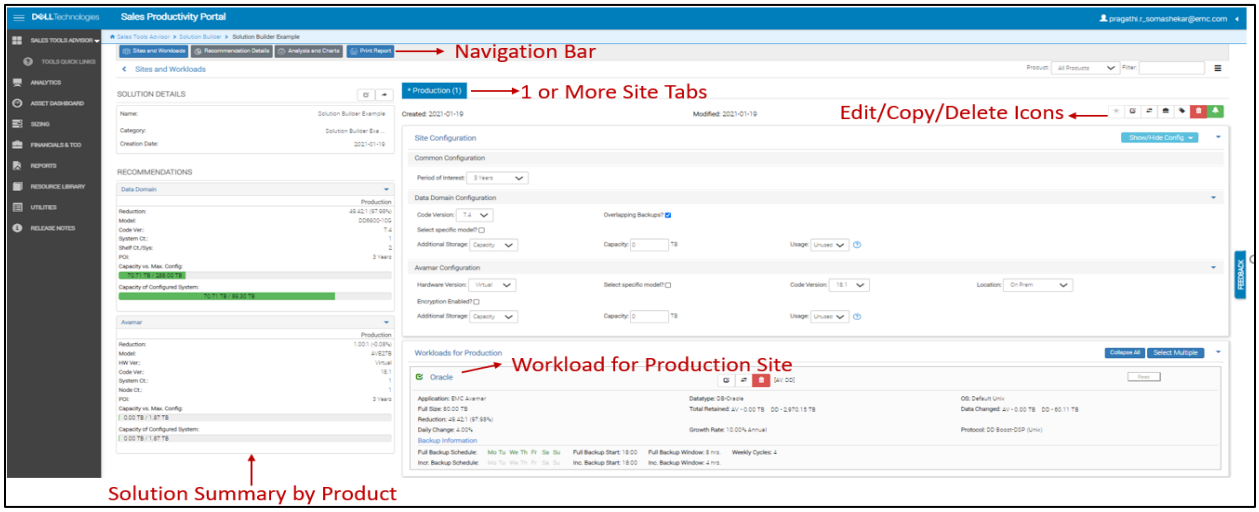


Figure 5 - Solution Builder Main Interface

Figure 6 - Solution Builder Navigation Bar

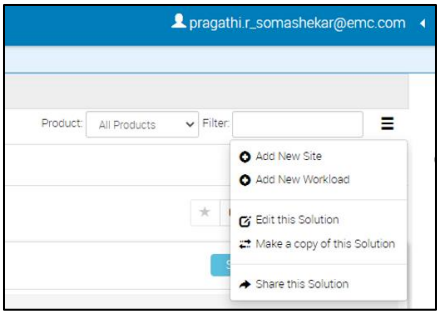


Figure 7 - Options under Hamburg Menu

Solution Builder Inputs

- Site configuration details (Network, Read, Replication).
- Details about various workloads to protect (FETB, growth rate, change rate, backup application, datatype, backup schedule)
- Backup Retention and location of replicated copies, Specify CloudTier & Cyber Recovery Retention if enabled.
- Duration of Backup and Replication Window
- Additional Workload Options:

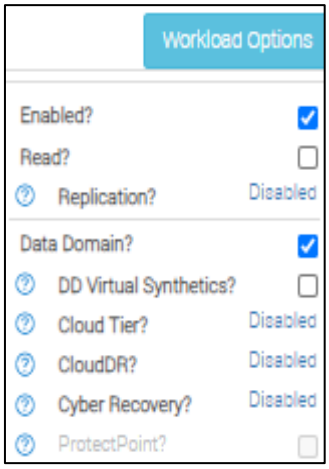


Figure 8 - Additional Workload Options

Solution Builder Output

- Recommended Models / Sizes of backup target devices (such as Data Domain, repository for Cloud Tier or Cloud DR).
- Analysis and charts: more in-depth details about the sizing solution, as well as charts and graphs that may better help visualize storage needs and requirements.
- Recommended replication bandwidth.

Solution Builder Workflow

1. Create a Solution.
2. Add at least one Site to the Solution (Option available in Hamburger menu).
3. Add at least one Workload to the Site (Option available in Hamburger menu).
4. Click the Recommendations icon to view the Solution Recommendations.

Multi-Edit Workloads

It is tedious and time-consuming to edit each and every workload in a site, in order to simply modify the annual growth rate from 10% to 15%, or to change the backup schedule, or to replace the relationships, i.e. to enable replication between Production and DR, or to enable Cloud tier between the Source site and the Cloud Target or to move workloads to Cyber Recovery vault. This task is very time consuming when there are n number of workloads in a site.

The above problem is solved by using the **Multi – Edit** Option available in Solution Builder.

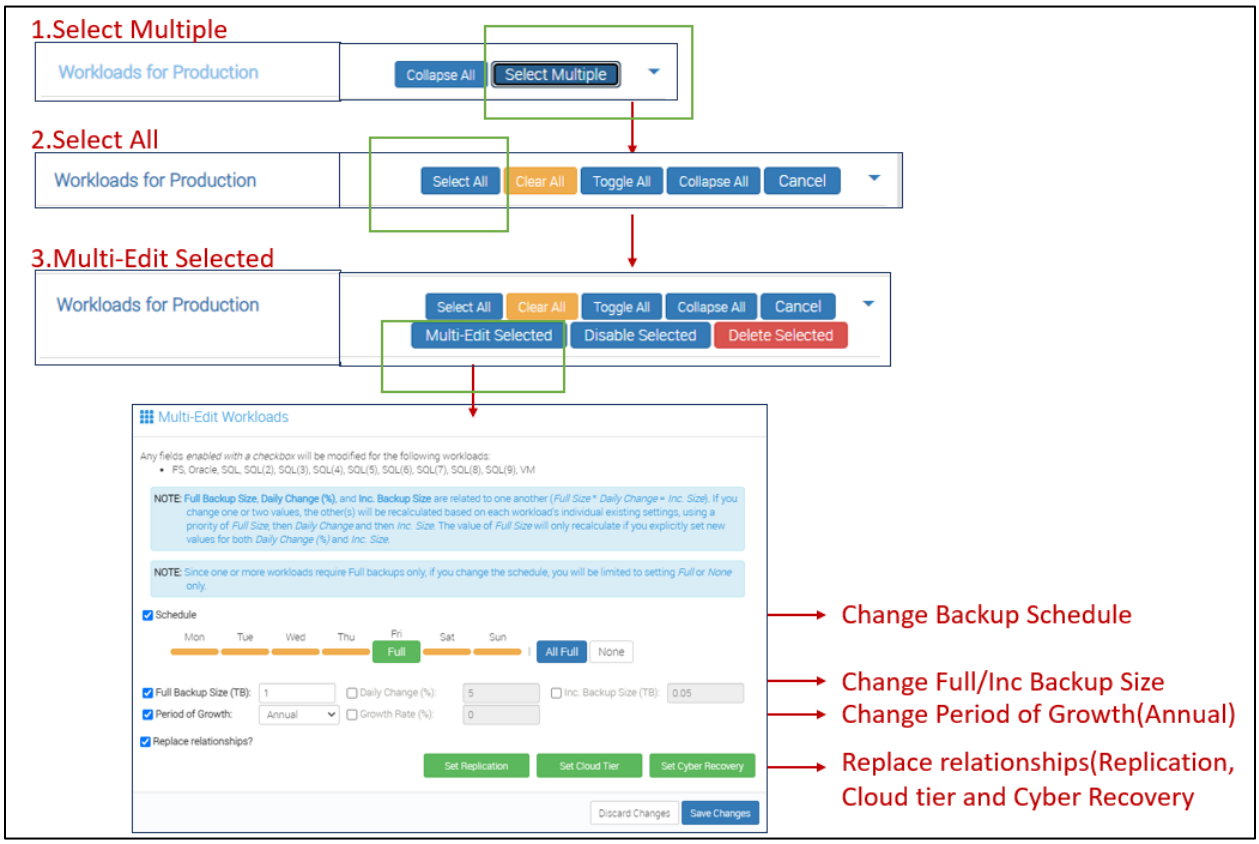


Figure 9 - Multi Edit Option Workflow

Example: To set up Replication context between Production and DR

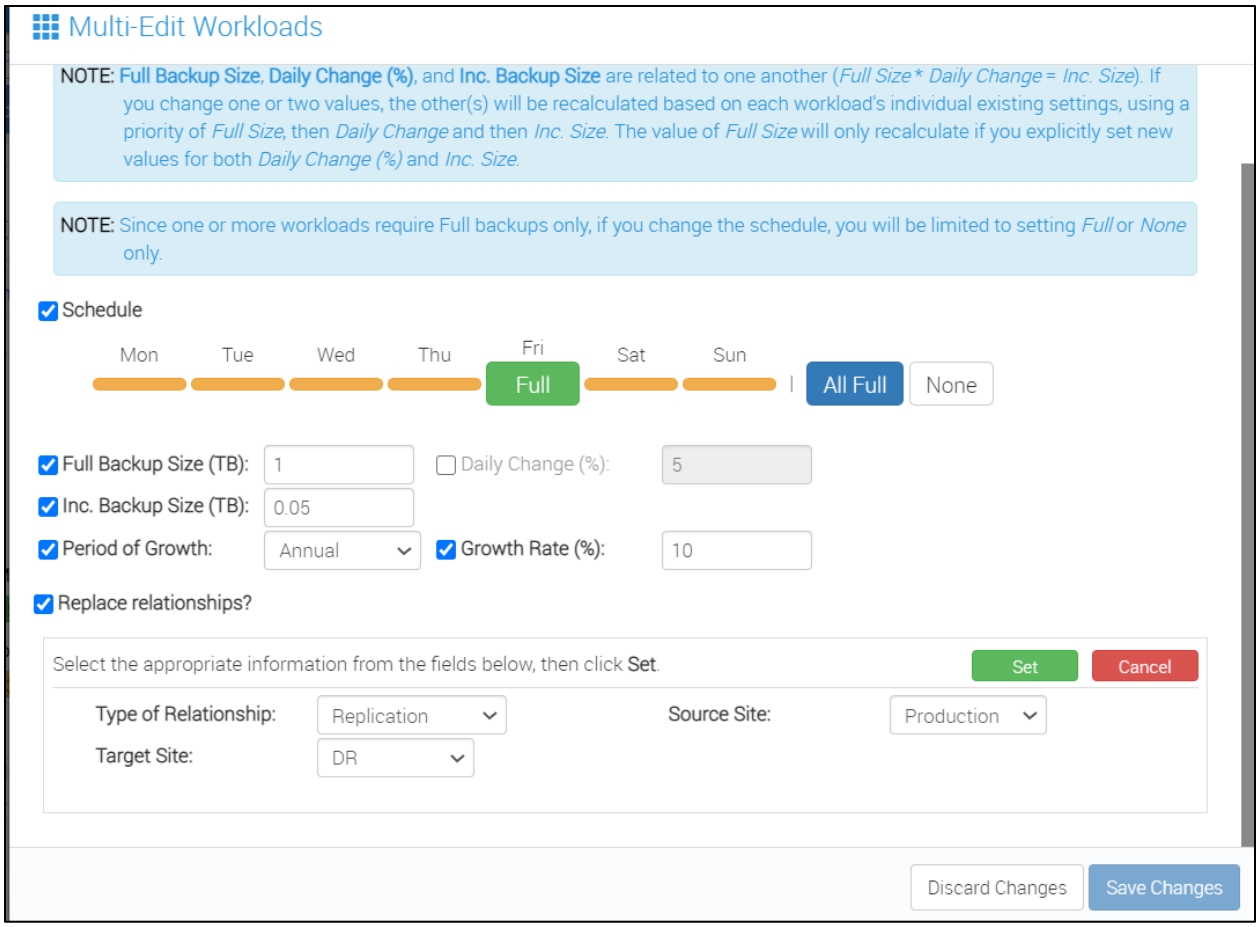


Figure 10 - To set up replication context using multi edit option

Solution Collaboration

- You can share your work with your colleagues by simply sending them the URL to the solution.
- Recipients can then review it in read-only mode, or copy it to their My Solutions list so that they may make changes without affecting your original work.
- A copy can be made by clicking the Copy to My Solutions button and then giving it a name.

Blue-Green Charts for Sizing

Blue Green charts corresponds with Solution Builder tool and is used as a tool to predict the yearly growth projection of the Data Domain system designed.

Blue-Green chart Workflow

1. **Dataset**: Enter the name of the workload being protected.
2. **Size (TB):** Enter the size of the workload.
3. **Retention**: Specify the retention period of each workload.
4. **Annual growth**: Specify the annual growth of each workload

5. **Data Protected** and **Space Required**: Capture the above values for **Data Protected** and **Space Required** of the Blue Green chart from the Solution Builder for the corresponding years.

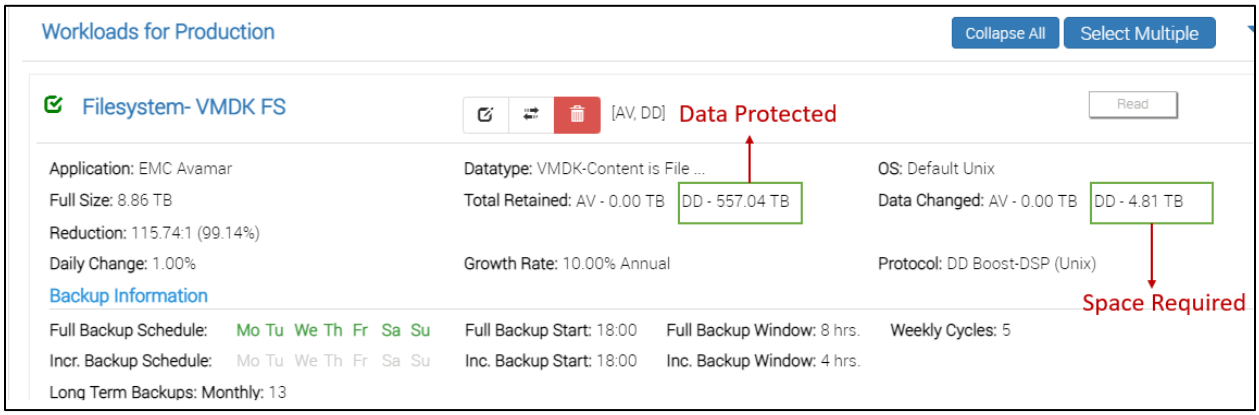


Figure 11 - To capture data protected and space required value from Solution Builder

Example: Assume a customer is looking for a 3-year solution and around 40 TB of Usable capacity. Considering growth rate, a Data Domain 6900 model is recommended. Now when we size the solution for this capacity and model type, the sizer recommends the solution mentioned in Figure 13 based on the assumptions shown in Figure 12.

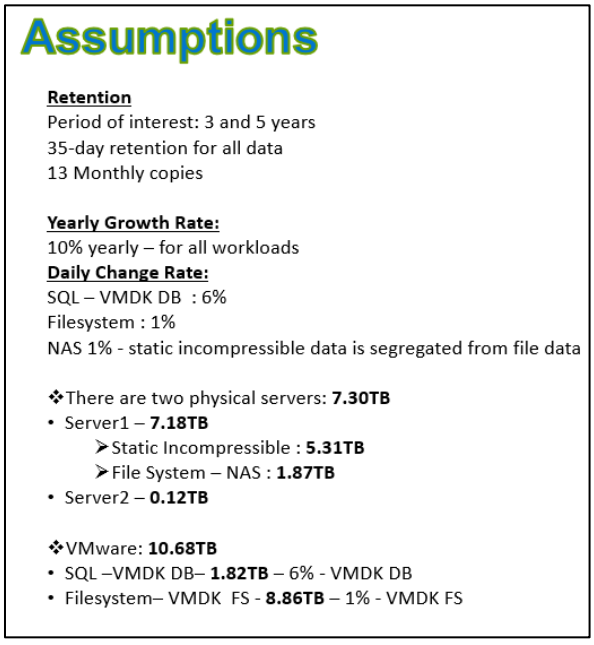


Figure 12 - Inputs for sizing (Both Blue-green and Solution Builder)

Solution Builder Output: In this case, Solution Builder recommends **DD6900** with **16.6TB Used/43.70TB Usable. (**Note that Solution Builder output includes 15% Buffer**)**

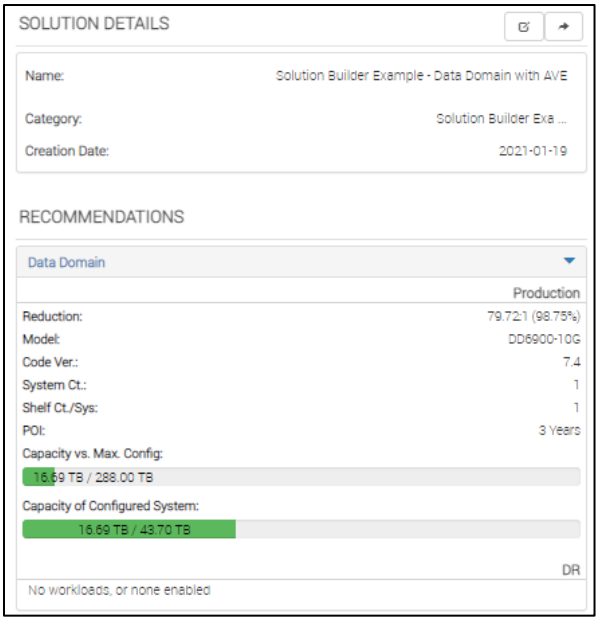


Figure 13 - Solution builder output for the above assumption (Figure 11)

Solution Builder URL for the Data Domain Sizing based on above assumptions:
https://salestools.isus.emc.com/#/sb/15cf39a958f87ae68f189ff8b0636f99

Blue Green Chart Output: The output obtained through Blue Green Chart is **14.2TB Used** (without any buffer %) and it matches the Solution Builder Output.

Solution Builder Output – 16.69TB Used

With 15% Buffer, 0.15TB x 16.69TB = 2.50TB

i.e.16.69TB – 2.50TB (Subtracting Solution Builder Output from the Buffer value) = **14.2 TB**, which is same as Blue Green Output at the end of 3 years.

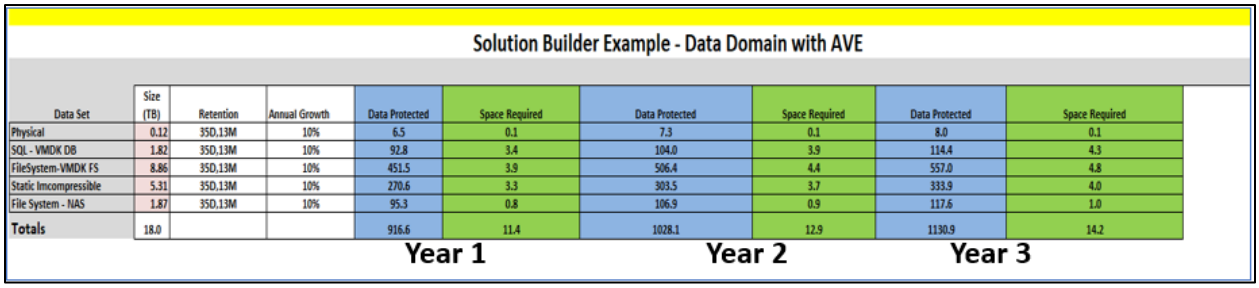


Figure 14 - Blue-green chart output for the assumption (Figure -11)

Data Protection Storage Capacity Projection graph

Figure 15 depicts the 3-year Data Protection Storage Capacity Projection graph; **Blue** represents the data we are protecting, and **Green** represents the data getting stored using Data Domain. The graph clearly shows that after the data reduction techniques (compression and deduplication) less data protection storage is required to protect large amount of workload data.

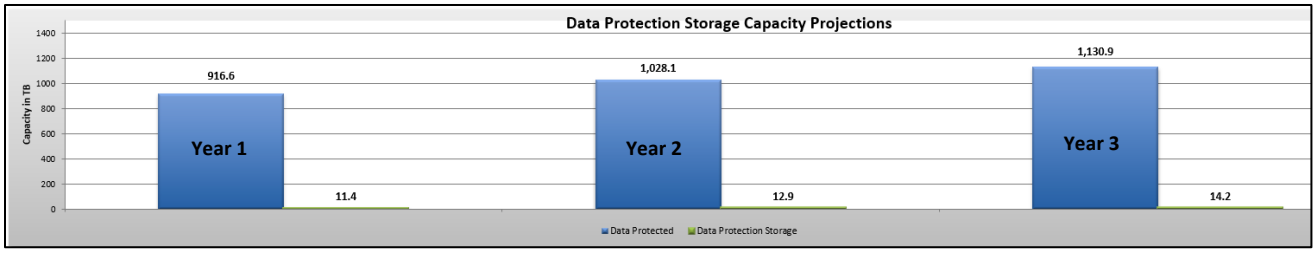


Figure 15 - Data Protection growth projection

Other Best Practices for Sizing

- Some requirements will not always be given, so you will need to make few assumptions.
- Always document all assumptions you will be making.
- Use DPS Integrated High IQ for Reduction Aggressiveness for DD to get better reduction values.
- Typically, the Daily change rate is 1-2 % for VM's, 3-5 % for Database, 1-2 % for NAS. However, there is no hard rule; it could be anything depending on the customer.
- Always check recommendation for results driven by Storage/Performance and modify your solution accordingly.
- File count/file size can significantly impact metadata storage required by backup application.
- Generally, a customer should have at least two copies of their backup data. (Don't forget Replication and Cyber Recovery).
- Customers generally have long-term as well as short-term retention requirements.
- Sizing is an Iterative exercise. It keeps changing!!

Sizing for accurate capacity and performance

Capacity and throughput planning are two critical aspects which determines how backup efficiency. Inadequate capacity and throughput planning can result in backup failure, backup lags or even data loss. Various factors need to be considered during capacity and performance planning, including:

Capacity Planning

The goal of capacity planning is to size a solution using the appropriate model and configuration which can accommodate the required data backups over a given period of time.

The first step in capacity calculation is to capture the following sizing information:

- **Data Size**: refers to amount of data in TB to be sized.
- **Data Type**: refers to the type of data e.g., Filesystem, database, VMDK, etc.
- **Retention Policy**: determines how long the data needs to be retained.
- **Schedule**: refers to the frequency of backups.
- **Period of Interest**: POI refers to period in years upfront for which you are sizing the solution, e.g. Solution Builder provides POI options from 1 month to 10 Years.
- **Growth rate**: refers to average data growth year over year.
- **Change rate**: refers to change in data, i.e. average percent of new data being backed up, e.g. Database data type has higher change rate compared to filesystem.

Using information collected about the backup system, you need to calculate capacity needed by understanding the amount of data (data size) to be backed up, the types of data, the size of a full (complete) backup, the number of copies of the data backed up, and the expected data reduction rates (deduplication).

Please find the attached link to a common DPS questionnaire to collect your customer's environment details for sizing : DPS_Questionnaire

Factors affecting capacity

1. Data Size
2. Data type
3. Data reduction rate (Deduplication)
4. Change rate
5. Retention

NOTE:

- Types of challenging (deduplication-unfriendly) datatypes include:
  - Compressed (multimedia, .mp3, .zip, and .jpg)
  - Encrypted data
- Data reduction factors depend on the type of data being backed up.
- Retention policies greatly determine the deduplication rate of a Data Domain system. Simple logic being the longer data is retained, the greater the data reduction that can be realized. For instance, a backup schedule where retained data is repeatedly replaced with new data results in little data reduction.

Data Domain system model capacity

Data Domain enables organizations to protect, manage and recover data at scale across their diverse environments. Dell Data Protection offers a comprehensive Data Domain portfolio to fit the requirements of organization's of any size and use case.

Comprehensive DD series portfolio

	DDVE - 96TB	DD3300	DD6900	DD9400	DD9900
Backup Ingest (w/DD Boost)	Up to 11.2TB/hr	Up to 7.0TB/hr	Up to 33TB/hr	Up to 57TB/hr	Up to 94TB/hr
Logical Capacity (w/Active Tier)	Up to 4.8PB	Up to 1.6PB	Up to 18.7PB	Up to 49.9PB	Up to 97.5PB
Usable Capacity (w/Active Tier)	1TB-96TB	4TB-32TB	24TB-288TB	192TB-768TB	576TB-1.5PB

Figure 16 - Data Domain portfolio with Capacity and throughput details

Depending upon the capacity required, choose an appropriate Data Domain model and configuration. Best practice is to be conservative in capacity planning. Therefore, apply your requirements against conservative ratings of the Data Domain system.

Keep in mind to always use:

80% of model capacity i.e. factor a 20% buffer for capacity.

Capacity % = (Required Capacity/Max usable system capacity) *100

Example: Suppose a customer estimates they need 150 FETB to data for the next 3 years. Which Data Domain model would be suitable?

Assuming 10% YoY growth on 180 TB at the end of 3rd year, about 200 TB of usable storage is needed.

Starting off, consider DD6900.

Max system usable capacity = 288 TB

Required Capacity = 200 TB

Therefore, **Capacity%= (200/288) *100 = 69%**

If the capacity or throughput for a particular model does not provide at least a 20% buffer, calculate the capacity and throughput for a Data Domain model of the next higher capacity. For example, if the capacity calculation for a DD6300 yields a capacity percentage of 91%, only a 9% buffer is available. Thus, look at the DD6800 next to calculate its capacity.

Keep in mind sometimes one model provides adequate capacity, but does not provide enough throughput, or the other way around. The model selection must accommodate both capacity and throughput requirements with an appropriate buffer. With that in mind, let's explore factors affecting throughput performance and how to size based on throughput.

Performance Planning

While capacity is one aspect of sizing it is important not to neglect the throughput required for efficient backups. The below formula gives you the required throughput:

Required throughput = Largest backup divided / backup window time

For instance, for an unrestricted network of 1GB, with maximum bandwidth available (having a theoretical 260 GB per hour throughput), would take less than 1 hour to complete a backup.

Note that if the network were sharing throughput resources during the backup window time, the amount of time required to complete the backup would increase considerably.

It is also important to note the effective throughput is a result of both the Data Domain system and the network on which it runs. Hence, both these points in data transfer determine whether the required speeds are reliably feasible. Feasibility can be assessed by running network testing software such as iperf.

Maximum throughput for each Data Domain model depends mostly on the number and speed capability of the network interfaces being used to transfer data. Some Data Domain systems have more and faster processors so they can process incoming data faster.

The number of network streams you may expect to use depends on your hardware model. To learn specific maximum supported stream counts, see the System Guide of that specific Data Domain model.

Factors affecting Data Domain Performance

Both external and internal factors affect Data Domain system performance in backup environments.

- **External Factors** in the backup environment that often gate how fast data is sent to the Data Domain system. External factor bottlenecks do not affect the potential throughput of the Data Domain system.

- **Internal Factors** reduce potential throughput of the Data Domain system. Internal factor bottlenecks require that specified values are addressed for potential sustained performance of the Data Domain system. Internal factors include:
    - Simultaneous streams
    - Garbage collection
    - Initial dataset backup speeds
    - Compression
    - High replication load
    - RAID re-build

The number of **simultaneous streams** in a Data Domain system significantly effects its potential throughput. As shown in Figure 16, each Data Domain system model has a range of simultaneous stream counts for optimal throughput and peak efficiency.

Data Domain systems are designed such that performance does not drop below 85% of peak throughput when more streams are used (up to the maximum number supported for the model and protocol). It is not recommended to run with more streams than the maximum supported values, as it is not tested and can reduce system performance.

Garbage Collection (GC) refers to reclaiming space that unreferenced data segments use; also known as cleaning. Data segments become unreferenced due to backup application deletion policies. The cleaning process impacts Data Domain performance.

Other **internal factors** include Initial dataset backup speeds, compression, RAID rebuilds, and high replication load. Ensure that you follow the most current best practice for each protocol.

The **number of disks** in a Data Domain system is an important factor for the level of performance the system can achieve.

Data Domain system model for Performance

Similar to choosing an appropriate model to meet capacity requirements, let's factor in performance requirement to choose an appropriate Data Domain model.

Apply your requirements against conservative ratings (not the maximums) of a Data Domain system. Allow for a minimum 20% buffer in both capacity and throughput requirements:

Throughput % = (Required throughput / maximum throughput of a particular model) * 100

With capacity, performance, and buffer all considered, this is a best practice for selecting a Data Domain model that is being implemented.

Example: A customer estimates that they require 275-TB usable storage for backups over the next 5 years. They require at least 15 TB/hour throughput to ensure that all data is backed up within their backup window. Which model of Data Domain suits their requirement?

The customer could use the DD6900 if both DD Boost and Cloud Tier are used. Otherwise the DD9400 would be the better choice both in terms of capacity and performance.

Conclusion

In any backup environment, it is critical to plan for adequate capacity and throughput. Planning ensures your backups complete within the time required and are securely retained for the needed times. The main goal in capacity planning is to design your system with a model and configuration that can store the required backup data for the specified retention period. To size an optimal solution, it is important to calculate the capacity and performance required. The methodology documented in this article will give sales and presales personnel an idea of how to quickly calculate capacity and performance metrics based on collected customer workload details, understand the metrics that impact performance and capacity of the backup solution, and how each metric is relatively dependent which ultimately helps drive their sales /presales opportunities.

References

https://salestools.isus.emc.com/

http://asup.corp.emc.com/ddasup

https://pdi.brs.lab.emc.com/Elysium/

https://inside.dell.com/people/fifern/blog/2019/11/10/fifer-scripts-data-domain-asup-parser

https://dell.sabacloud.com/Saba/Web_spf/PRODTNT091/app/shared;spf-url=common%2Flearningeventdetail%2Fcurra000000000009651