IBM® Storage

IBM Storage for AI Solutions for Splunk Enterprise (Traditional Deployment)

Architectural Guide

Contents

About this document..

Executive summary 1

Scope 1

Prerequisites.. 1

Getting started: Splunk Enterprise Data Solutions building blocks using IBM Storage........... 2

- Splunk Enterprise Application 2
- Splunk Enterprise deployment architectures... 2
- How Splunk Stored the data:.. 4
- IBM Storage Solutions for Splunk Enterprise 4
- Solution LAB Validation 8
- Storage Sizing Guidelines.. 12

Summary 13

Notices... 9

Trademarks 9

Terms and conditions for product documentation 11

- Applicability 12
- Commercial use 13
- Rights 14
- Privacy policy considerations 14

About this document

This document is intended to facilitate the deployment of the scalable Splunk Enterprise data architecture using IBM All Flash Array System, IBM Elastic Storage System (ESS) and IBM Spectrum Scale. To complete the tasks it describes, you must understand IBM All Flash Array, IBM Spectrum Scale system and IBM Elastic Storage System.

The information in this document is distributed on an *as is* basis without any warranty that is either expressed or implied. Support assistance for the use of this material is limited to situations where IBM Flash System, IBM Spectrum Scale are supported and entitled, and where the issues are specific to a blueprint implementation.

Executive Summary

Splunk® Enterprise has become a mission critical application. Organizations are using the Splunk Enterprise Application to gain insights from the logs and machine data and becoming one of the critical applications in the environment. IBM is the leading storage systems provider and offers the NVMe based all flash arrays for the high-performance tier and enterprise file storage systems based on IBM Spectrum Scale technologies for holding the larger data sets as the data life cycle management. This solution blueprint provides a framework for building the Splunk enterprise environment using the IBM Storage Portfolio.

Scope

This blueprint guide provides the following information:

- A solutions architecture and related solution configuration best practices, with the following essential software components:
  - IBM All Flash Array System
  - IBM Elastic Storage System
  - IBM Spectrum Scale
- Detailed technical configuration steps for building an end-to-end solution in the hybrid cloud environment

This technical report does not make the following changes:

- Provide performance analysis from a user perspective
- Replace any official manuals and documents issued by IBM

Prerequisites

This technical paper assumes basic knowledge of the following prerequisites:

- IBM All Flash Array Systems
- IBM Elastic Storage System
- IBM Spectrum Scale
- Splunk Enterprise Software and Solutions

Getting started: Splunk Enterprise Data Solutions using IBM Storage

Globally, Datacenters leverage different systems to deliver business services. It is an extremely complex effort to efficiently manage the abundance of devices, deployed in a typical datacenter. Several devices experience outages, performance issues, or missed SLA's, on a daily basis. To ensure high performance and availability in the enterprise, IT administration teams waste valuable resources accessing several management consoles, and manually run home-grown scripts to serially trace the valuable data they need from failed devices. This is machine data, a form of Big-Data. Splunk Enterprise provides easy visibility, reporting and search across all IT systems and infrastructure in the enterprise. It delivers strong machine-data governance, with comprehensive controls for data security, retention and integrity.

IBM provides a high very performance, scalable and multi-tier storage platform for the Splunk enterprise, to mine this data operationally, and provide deeper insight into the workings of any Data Center, real-time. IBM solution enables easy mining of data stored in various configuration files, log files, network ports, databases, IO ports, trace files, application code and scripts, file systems, event logs and many other sources.

IBM realize that there are different types of clients in the enterprises and mid-markets, and they all have varying sizes of data requirements. Within the IBM portfolio, a choice of performance proven storage systems offers unique storage capabilities with a data-centric view of resources and shared data repositories prevalent in local, distributed, or replicated datacenters. These storage systems are identified, as follows:

1. IBM Flash system
2. IBM Spectrum Scale
3. IBM Elastic Storage System

Splunk Enterprise Application

Splunk Enterprise is a software product that enables you to search, analyze, and visualize the data gathered from the components of your IT infrastructure or business. Splunk Enterprise takes in data from websites, applications, sensors, devices, and so on. After you define the data source, Splunk Enterprise indexes the data stream and parses it into a series of individual events that you can view and search.

Splunk Enterprise indexes data from the servers, applications, databases, network devices, and virtual machines that make up your IT infrastructure. As long as the machine that generates the data is a part of your network, Splunk Enterprise can collect the data from anywhere, whether the data is local, remote, or in the cloud.

Splunk Enterprise performs three main functions as it processes data:

1. It ingests data from files, the network, or other sources.
2. It parses and indexes the data.
3. It runs searches on the indexed data.

Splunk Enterprise Deployment Architectures

Depending on your needs, you can deploy Splunk Enterprise as a single instance, or you can create deployments that span multiple instances, ranging from just a few to hundreds or even thousands of instances.

Single-instance deployments

In small deployments, one instance of Splunk Enterprise handles all aspects of processing data, from input through indexing to search. A single-instance deployment can be useful for testing and evaluation purposes and might serve the needs of department-sized environments.

Distributed deployments

To support larger environments where data originates on many machines, where you need to process large volumes of data, or where many users need to search the data, you can scale the deployment by distributing Splunk Enterprise instances across multiple machines. This is known as a "distributed deployment".

In a typical distributed deployment, each Splunk Enterprise instance performs a specialized task and resides on one of three processing tiers corresponding to the main processing functions:

- Data input tier (forwarder)
- Indexing tier (Indexer)
- Search management tier (Search head)

These specialized instances are known as "components". Below is the basic simple architecture of the Splunk deployment model.

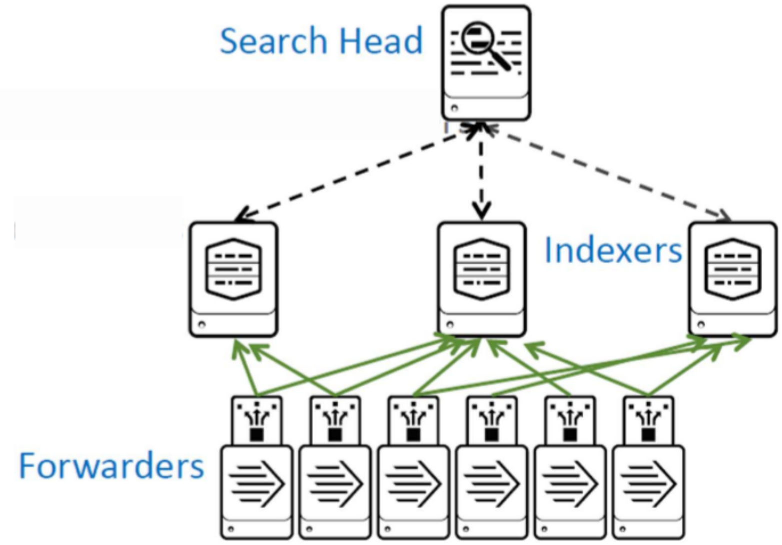


Figure1: Splunk distributed architecture

- **Forwarder:** A Splunk Forwarder monitors files, detects file changes, listens to network ports, and executes data gathering functions. A forwarder consumes data and then forwards the data onwards, usually to an indexer(s). Forwarders usually require minimal resources, allowing them to reside lightly on the machine generating the data.
- **Indexer:** It accepts new data (from forwarders), adds it to an index and compresses it on disk. An indexer indexes incoming data that it usually receives from a group of forwarders. The indexer transforms the data into events and stores the events in an index. The indexer also searches the indexed data in response to search requests from a search head. To ensure high data availability and protect against data loss, or just to simplify the management of multiple indexers, you can deploy multiple indexers in indexer clusters.
- **Search-Head:** It is the webserver and application interpreting engine that provides the primary, web-based user interface for reporting search results. Since most of the data interpretation happens as-needed at search time, the role of the search head is to translate user and app requests into actionable searches for its indexer(s) and display the results. It leverages REST protocols to communicate with the Indexing server at port 8000, to execute look-ups into the indexed data streams. To ensure high availability and simplify horizontal scaling, you can deploy multiple search heads in search head clusters.

  Buckets: Splunk Enterprise stores indexed data in buckets, which are directories containing files of data. An index typically consists of many buckets. A complete cluster maintains replication factor number of copies of each bucket, with each copy residing on a separate peer node. The bucket copies are either searchable or non-searchable. A complete cluster also has search factor number of searchable copies of each bucket.

How Splunk stores data

An index typically consists of many buckets, and the number of buckets grows as the index grows. As data continues to enter the system, the indexer creates new buckets to accommodate the increase in data. The number of buckets in an index can grow quite large, depending on how much data you're indexing and how long you retain the data. A bucket moves through several states as it ages:

- hot
- warm
- cold
- frozen
- thawed

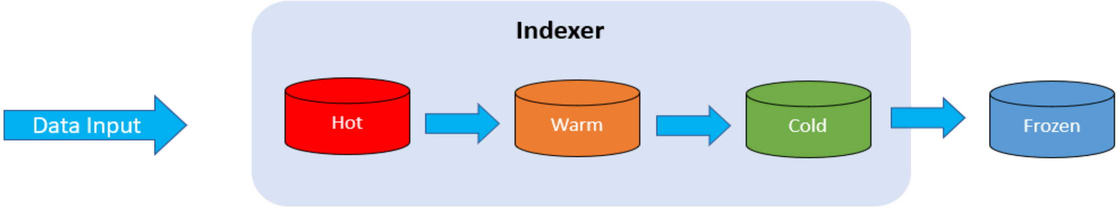


Figure 2: Splunk data storage buckets

Bucket type	Role	Searchable
Hot	New data is written to hot buckets. Open for read and write operations. One or more hot buckets for each index.	Yes
Warm	Data is rolled from hot bucket. Data is not actively written and will be read only at the index level. An index will have many warm buckets.	Yes
Cold	Buckets rolled from warm and moved to a different location. There are many cold buckets depending up on the retention period.	Yes
Frozen	Buckets rolled from cold. The indexer deletes frozen buckets, but you can choose to archive them first. Archived buckets can later be thawed	No
Thawed	Buckets restored from an archive. If you archive frozen buckets, you can later return them to the index by thawing them.	Yes

IBM Storage solutions for Splunk Enterprise

Splunk distributed architecture model allows organizations to design and scale the architecture as per the data ingest and search requirements. Medium to Large enterprise architecture allows a daily ingest data rate of 250 GB per Indexer server.

IBM Flash system for hot and warm buckets

Splunk Hot/Warm Buckets are actively written and being used during the search operations. Splunk recommends high performance block storage tier for hot and warm tiers. Volume of Ingest, Search types, application type and number of users will impact the performance and capacity characteristics of the hot and warm buckets. Hot buckets are active and there are multiple read/write activities performed on this tier. When a hot bucket fills and reaches the threshold limit, it becomes the warm bucket and stays on the same storage tier.

Local attached DAS configurations with SSD drives are good for smaller environment, however when the environment grows into hundreds of indexers, it results into capacity overheads and becomes complex from the management perspective.

When the Splunk environment grows and an external SAN storage becomes more reliable and cost-effective solution for hot/warm tiers. IBM offers NVMe based All Flash Arrays from the entry level to high end enterprise solutions based on the Spectrum Virtualize Platform.

IBM offers 3 different kind of NVMe bases solutions

- Storwize V5000 – Entry to Mid-Range
- Storwize V7000 – NVMe accelerated Mid-Range
- Flash9100 – NVMe for Enterprise customers.

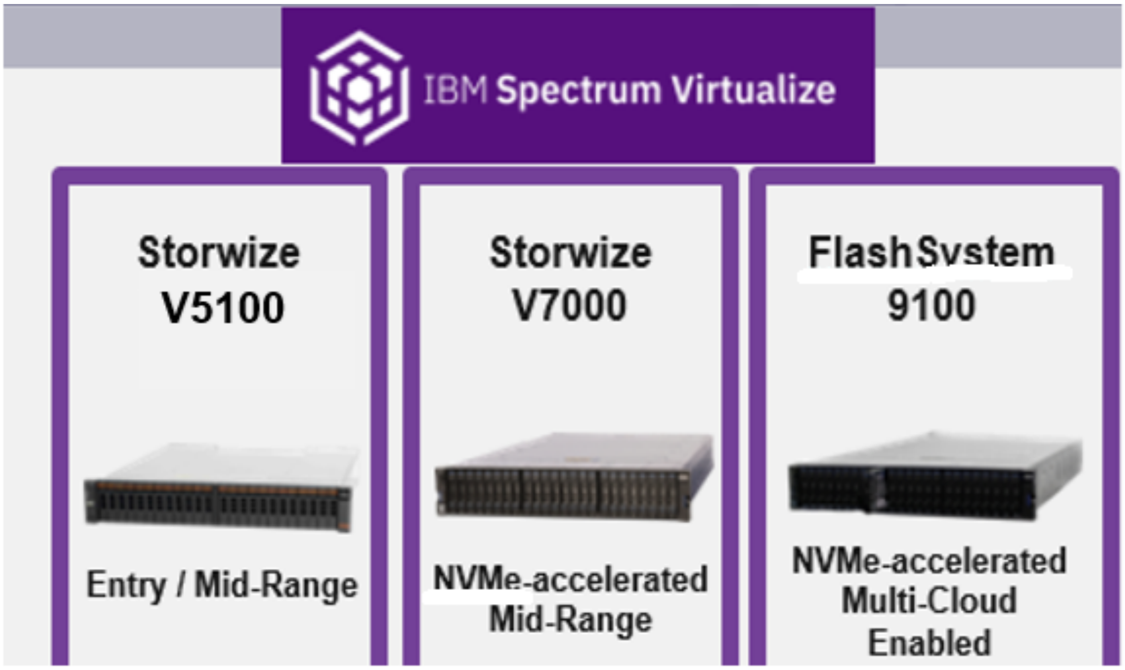


Figure 3: IBM NVMe based flash systems.

IBM Flash Systems combines the performance of flash and the Non-Volatile Memory Express (NVMe) protocol with the reliability and innovation of IBM FlashCore® technology and the rich feature set of IBM Spectrum Virtualize in one powerful new storage platform for your data-driven enterprise solution.

IBM Flash System provides the software-defined, modern data protection and multi-cloud capabilities of several members of the IBM Spectrum Storage™ family. IBM Spectrum Virtualize, the system foundation that provides a broad set of enterprise-class data services—such as dynamic tiering, replication, IBM FlashCopy® management, data mobility, transparent cloud tiering and high-performance data-at-rest encryption, among many others. The arrays also leverage innovative new data reduction pools (DRP) that incorporate deduplication and hardware-accelerated compression technology, as well as SCSI UNMAP support and all the thin-provisioning and data-efficiency features you expect from IBM Spectrum Virtualize-based storage to potentially reduce your capital and operating expenses. Additionally, IBM Spectrum Virtualization platform enables virtualization capabilities, which can be used to virtualize more than 440 IBM and non-IBM heterogeneous storage systems.

Benefits of IBM NVMe Arrays for Splunk

- Low latency
- up to 2:1 data reduction with low latencies.
- Data encryption
- Can virtualize more than 440 IBM and non-IBM storage systems.

IBM File Storage for Cold and Frozen bucket

As the data ages, Splunk moves the warm data to cold bucket as the low-cost storage tier. Typically, this will be a file-based storage and will use a Network Attached System (NAS) for holding the larger data sets as per the organization data retention policies. IBM offering in this area is Elastic Storage System (ESS) built using the IBM Spectrum Scale technologies.

IBM Spectrum Scale is a high-performance, highly available, clustered file system available on a variety of platforms including the public cloud service providers. It provides concurrent access to a single file system or set of file systems from multiple nodes. IBM Spectrum Scale has multiple data access points (via different protocols) where data in form of files and objects are directly accessible by end users as well as applications. Moreover, the product integrates with organization’s external directory services for different type of authentication as well as authorization.

Spectrum Scale file system can be built from a single disk or contain thousands of disks storing petabytes of data. Each file system can be accessible from all nodes within the cluster. There is no practical limit on the size of a file system. The architectural limit for a single file system is more than a yottabyte. Some Spectrum Scale customers use single file systems up to 18 PB in size, while others use file systems containing billions of files. Applications access files through standard Portable Operating System Interface (POSIX) file system interfaces or using standard industry-based standard protocols like NFS, SMB and Object interface. Spectrum Scale supports NFS v4 access control lists (ACLs) for the file access from the clients.

IBM Elastic Storage System is a modern implementation of software-defined storage, combining IBM Spectrum Scale software with IBM POWER8® processor-based I/O-intensive servers and dual-ported storage enclosures. IBM Spectrum Scale is the parallel file system at the heart of IBM Elastic Storage System. IBM Spectrum Scale scales system throughput as it grows while still providing a single namespace. This eliminates data silos, simplifies storage management and delivers high performance. By consolidating storage requirements across your organization onto IBM Elastic Storage System, you can reduce inefficiency, lower acquisition costs and support demanding workloads.

The capabilities of IBM Elastic Storage System include:

- **Declustered data:** IBM Spectrum Scale RAID distributes client data, redundancy information and spare space uniformly across disks. This distribution reduces the rebuild or disk-failure recovery process overhead compared to traditional RAID. Critical rebuilds of failed multi-terabyte drives full of data can be accomplished in minutes—rather than hours or even days when using traditional RAID technology.
- **Data redundancy:** IBM Spectrum Scale RAID supports highly reliable 2-fault-tolerant and 3fault-tolerant Reed-Solomon-based parity codes (erasure coding) as well as three-way and four-way replication.
- **Tuned performance:** Software-defined IBM Spectrum Scale RAID software, explicitly coupled with large memory cache in the IBM Power server, allows IBM ESS to mask the inefficiencies and long latency times of nearline-SAS drives with low latency flash storage, while still leveraging the high density of the drives themselves.
- **Simplified management:** The intuitive graphical user interface (GUI) for software and systems for management and monitoring of IBM ESS also integrates into IBM Spectrum Control.
- **Superior streaming performance:** The system can deliver more than 36 GB/s of sustained performance.
- **Scalability and extensibility with multi-site and cloud support:** Multiple deployment options for software-defined storage to scale in performance and capacity while still providing a single namespace. This means installations can start small and grow as data needs expand.

IBM offers two versions of the ESS product lines – Performance and Capacity models

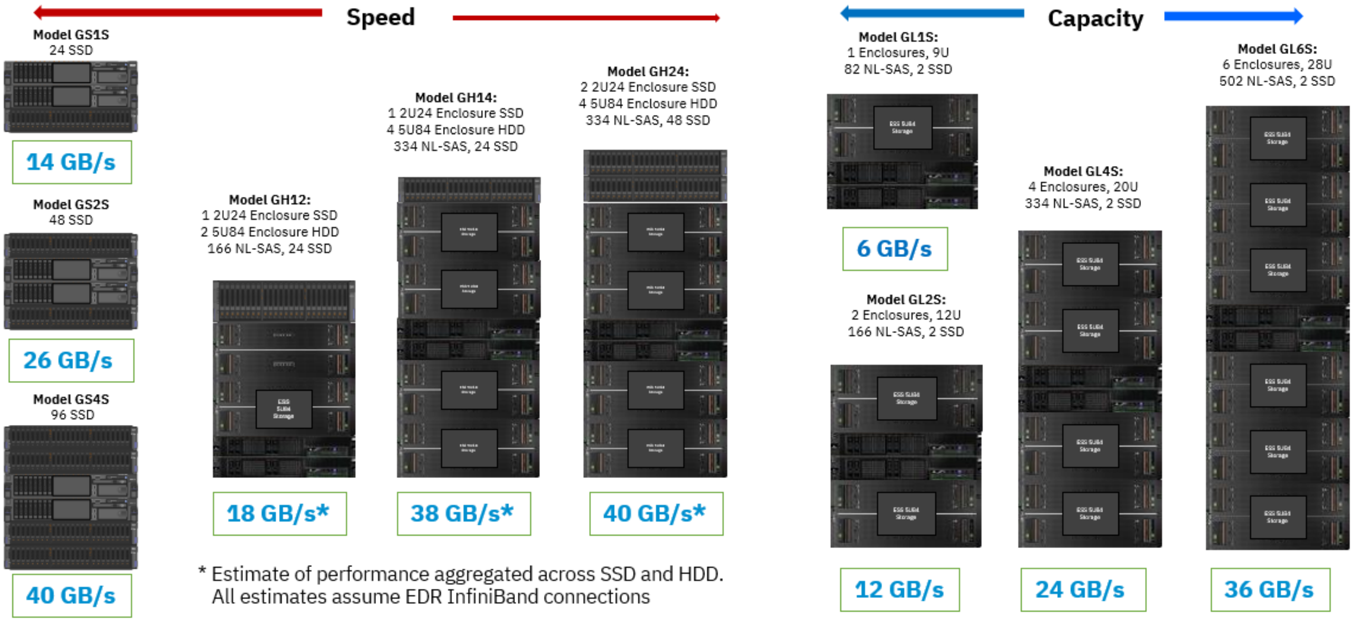


Figure 4: IBM Elastic Storage System family

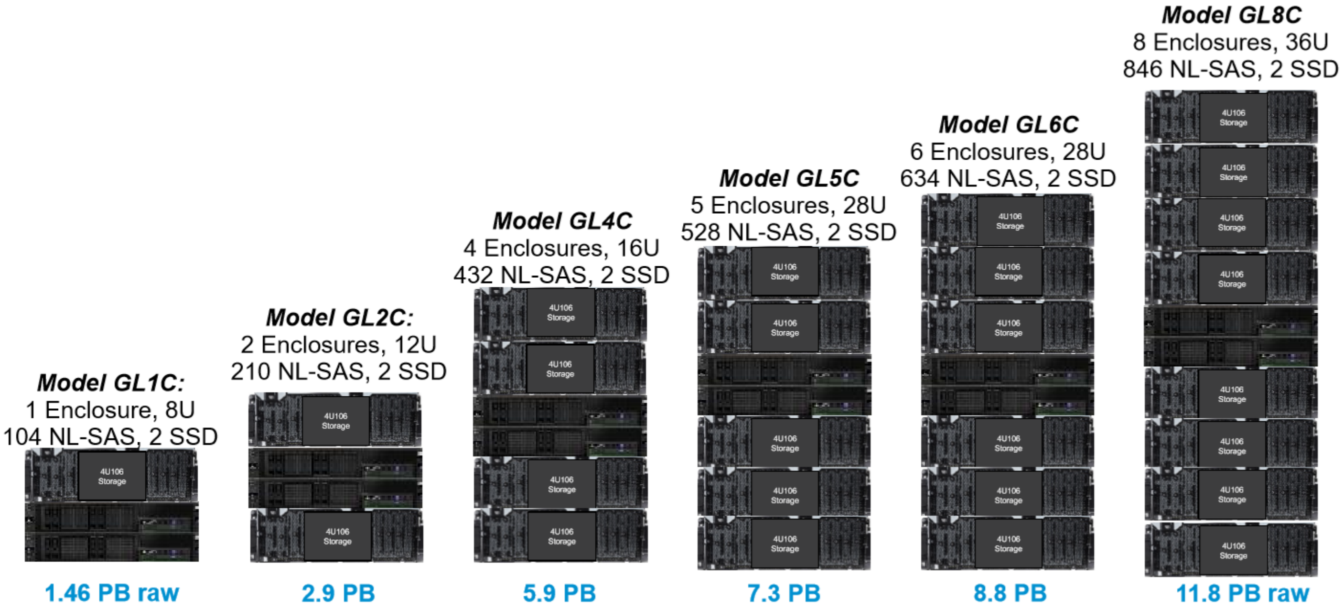


Figure 5: IBM Elastic Storage System capacity family – GLxC models

The benefits of Elastic Storage System include for cold and frozen buckets:

- Highest storage densities at the lowest costs in the industry today It provides upto 11.8 PB raw capacity in a single rack
- NFS support via protocol nodes.

IBM recommends NVMe based flash arrays for Hot/warm buckets and IBM Elastic Storage System via NFS protocol for Cold/Frozen buckets as mentioned in the below figure.

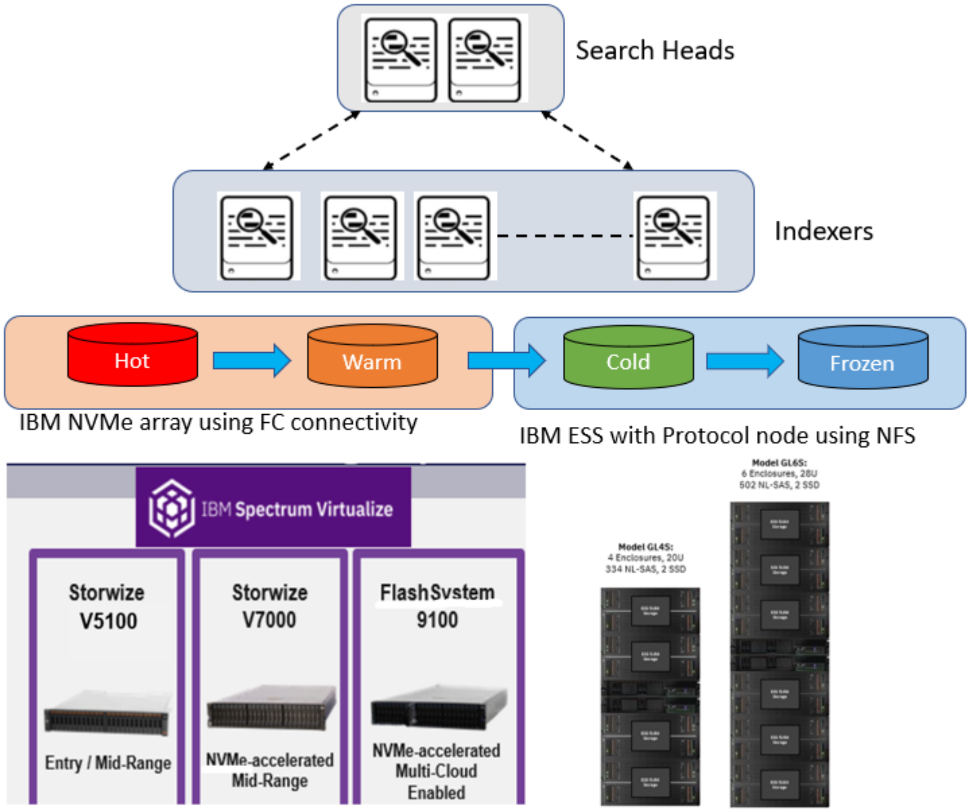


Figure 6: IBM Storage architecture for Splunk Enterprise

SOLUTION LAB VALIDATION

This section describes about the LAB installation, configuration and validation of the Splunk Enterprise with IBM Storage Portfolio. IBM flash system 9100 is configured for the hot and warm buckets and IBM ESS Storage with NFS protocol is used for Cold bucket. Splunk Enterprise version 7.3.1 is installed and configured in the LAB for the testing purpose. The Splunk Event Generator tool is used for simulating the events into the application.

LAB Configuration:

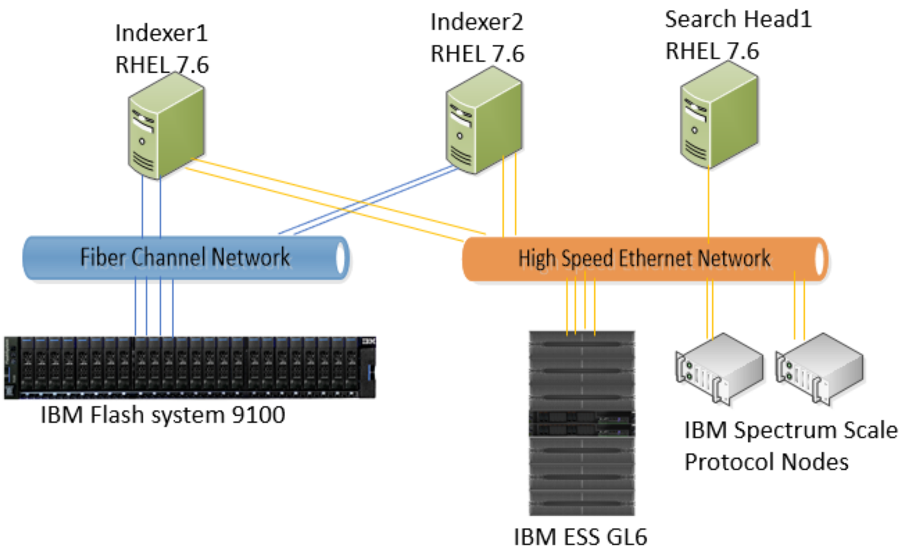
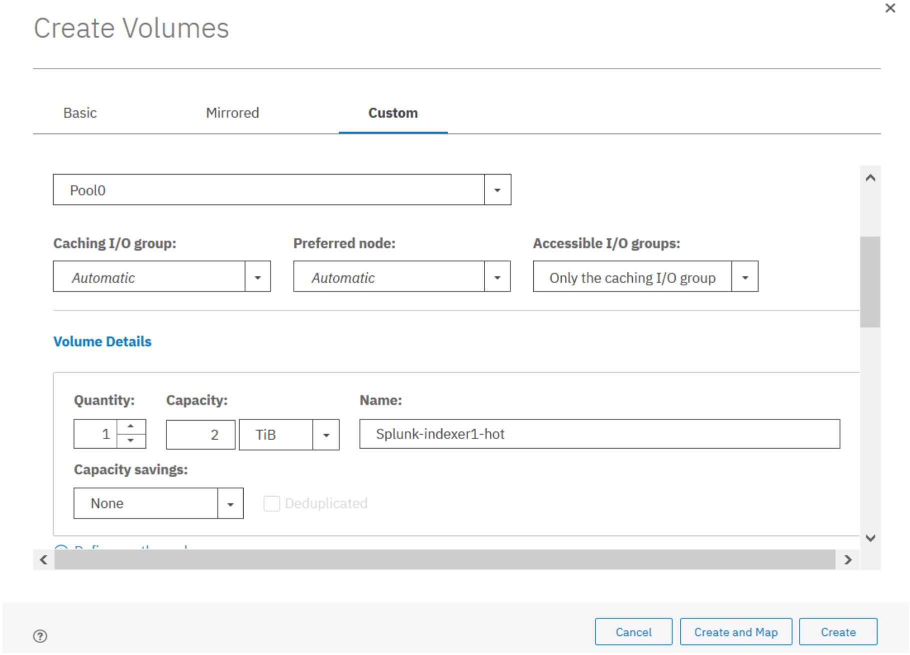


Figure 7: LAB configuration setup

- IBM Flash System 9100 is used for storing the hot and warm buckets.
- IBM ESS Storage with NFS protocol is used for Cold bucket.
- Splunk Enterprise version 7.3.1 is installed on Intel servers running with RHEL 7.6 OS.
- The Splunk Event Generator tool is used for simulating the events into the application
- Index Server connected with 16 Gig FC adapters to the FS9100 system
- I0 Gig Ethernet is used for connecting Indexers with ESS Protocol nodes.

1. Flash 9100 system configuration

a. Create the vdisk using the FS9100 console and assign it to the index servers as shown below



Create Volumes

Pool0

Caching I/O group: Automatic

Preferred node: Automatic

Accessible I/O groups: Only the caching I/O group

Volume Details

Quantity: 1

Capacity: 2 TiB

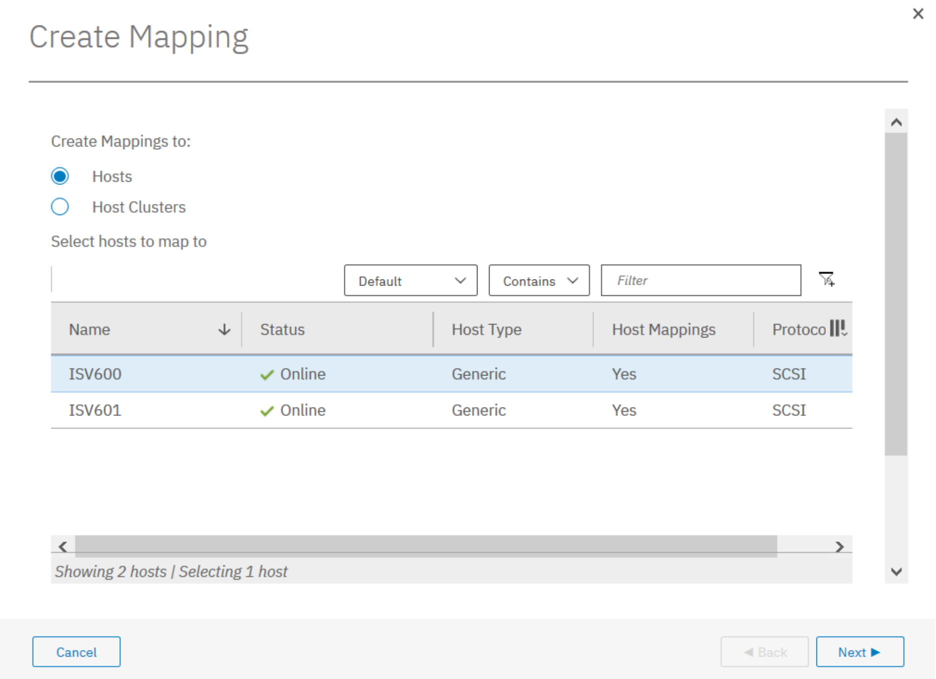
Name: Splunk-indexer1-hot

Capacity savings: None

□ Deduplicated

Figure 8: Vdisk creation using FS9100 GUI console

b. Map the vdisk to Indexer servers for storing the hot/warm data.


Create Mapping

Create Mappings to:

● Hosts

○ Host Clusters

Select hosts to map to

Name	Status	Host Type	Host Mappings	Protoco
ISV600	✓ Online	Generic	Yes	SCSI
ISV601	✓ Online	Generic	Yes	SCSI

Showing 2 hosts | Selecting 1 host

Figure 9: Hosting mapping for IO access

c. Create the file system and mount at the Indexer server.

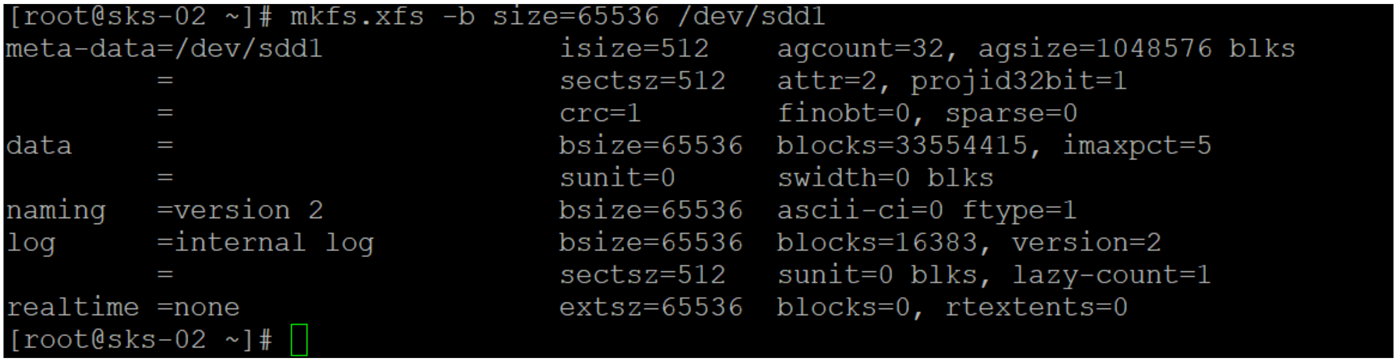

```
[root@sks-02 ~]# mkfs.xfs -b size=65536 /dev/sdd1
meta-data=/dev/sdd1              isize=512    agcount=32, agsize=1048576 blks
         =                       sectsz=512   attr=2, projid32bit=1
         =                       crc=1        finobt=0, sparse=0
data     =                       bsize=65536  blocks=33554415, imaxpct=5
         =                       sunit=0      swidth=0 blks
naming   =version 2              bsize=65536  ascii-ci=0 ftype=1
log      =internal log           bsize=65536  blocks=16383, version=2
         =                       sectsz=512   sunit=0 blks, lazy-count=1
realtime =none                   extsz=65536  blocks=0, rtextents=0
[root@sks-02 ~]#
```

Figure 10: xfs file system creation on the Indexer server

d. Mount the file system for the Splunk application usage on the Indexer server.

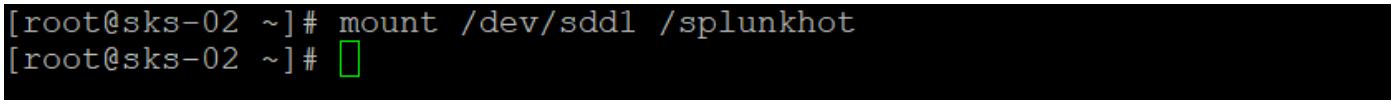

```
[root@sks-02 ~]# mount /dev/sdd1 /splunkhot
[root@sks-02 ~]#
```

Figure11: Mounting file system for hot/warm data usage on Indexer

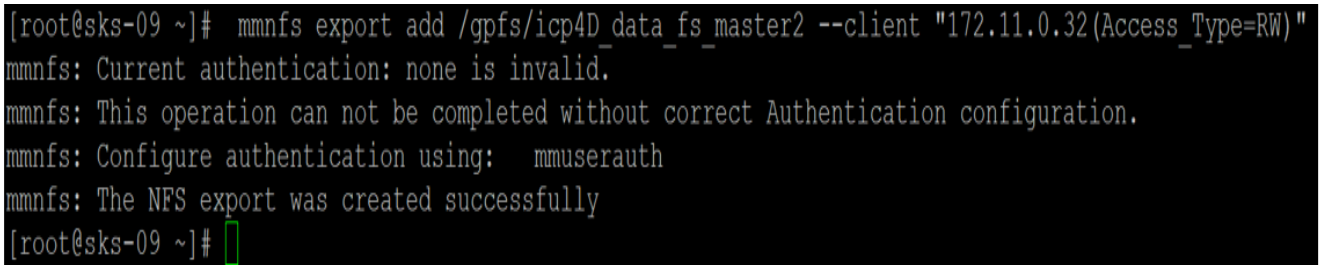
2. Configuring the IBM Elastic Storage System for Cold and Frozen buckets:

I. Splunk Indexer servers access the ESS storage via NFS protocol.

II. Minimum two protocol nodes are recommended for basic failover capabilities. Additional protocol nodes can be configured as per the scalability and performance characteristics of the Splunk workload.

III. Configure the Cluster Export Service (CES) services on the protocol node as mentioned in the below documents for NFS access. https://www.ibm.com/support/knowledgecenter/en/STXKQY_5.0.3/com.ibm.spectrum.scale.v5r03.doc/bl1adm_cesproconfig.htm

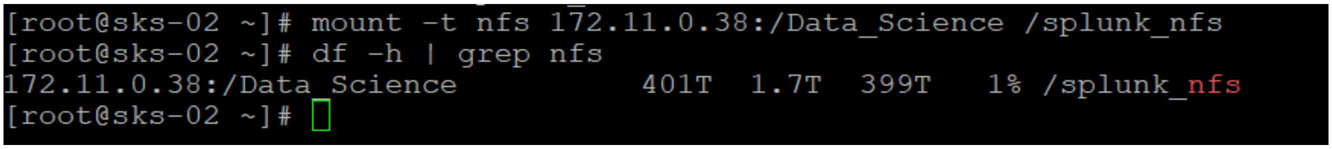
IV. Create the NFS export and share it with the Splunk Indexer server.



```
[root@sks-09 ~]#  mmnfs export add /gpfs/icp4D_data_fs_master2 --client "172.11.0.32(Access_Type=RW)"
mmnfs: Current authentication: none is invalid.
mmnfs: This operation can not be completed without correct Authentication configuration.
mmnfs: Configure authentication using:   mmuserauth
mmnfs: The NFS export was created successfully
[root@sks-09 ~]#
```

Figure12: Exporting Scale file system for NFS access

V. Mount the nfs export share on the Splunk the Index server



```
[root@sks-02 ~]# mount -t nfs 172.11.0.38:/Data_Science /splunk_nfs
[root@sks-02 ~]# df -h | grep nfs
172.11.0.38:/Data_Science          401T  1.7T  399T   1% /splunk_nfs
[root@sks-02 ~]#
```

Figure13: Exporting Scale file system for NFS access

3. Configuring the storage buckets at the Splunk Enterprise application

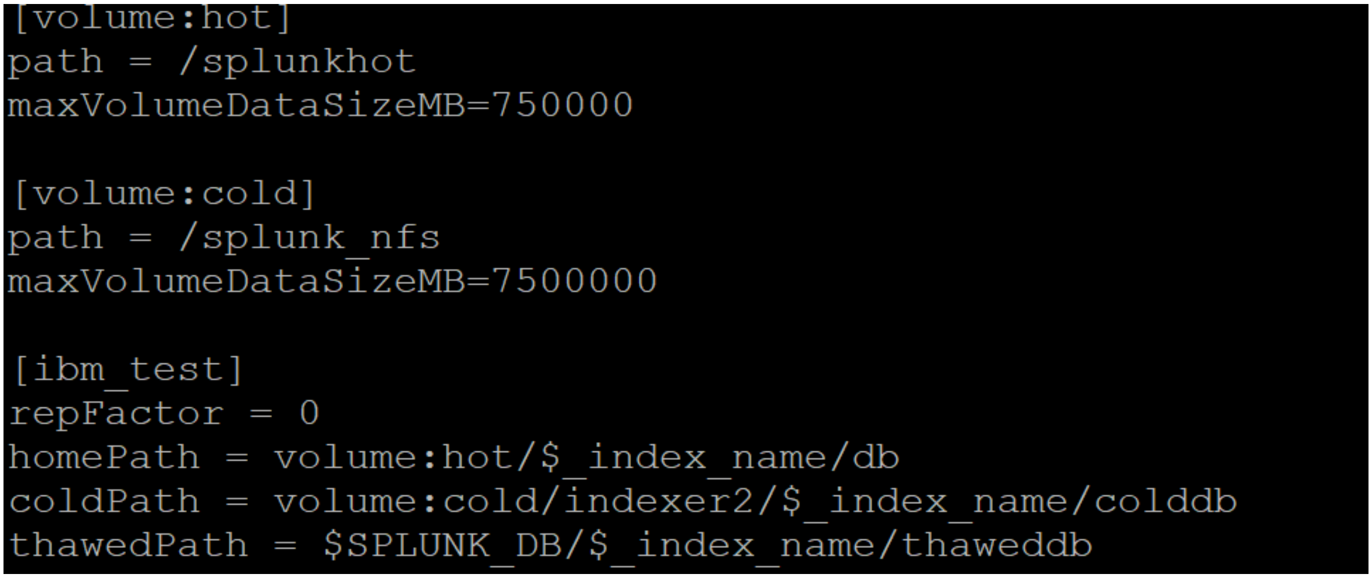
Splunk storage bucket configurations ae controlled via the parameters configured in the file “indexes.conf”

For a standalone indexer, edit the following file:

$SPLUNK_HOME/etc/system/local/indexes.conf.

For a cluster of indexers, edit the following file on the master node:

$SPLUNK_HOME/etc/master-apps/_cluster/local/indexes.conf



```
[volume:hot]
path = /splunkhot
maxVolumeDataSizeMB=750000

[volume:cold]
path = /splunk_nfs
maxVolumeDataSizeMB=7500000

[ibm_test]
repFactor = 0
homePath = volume:hot/$_index_name/db
coldPath = volume:cold/indexer2/$_index_name/colddb
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
```

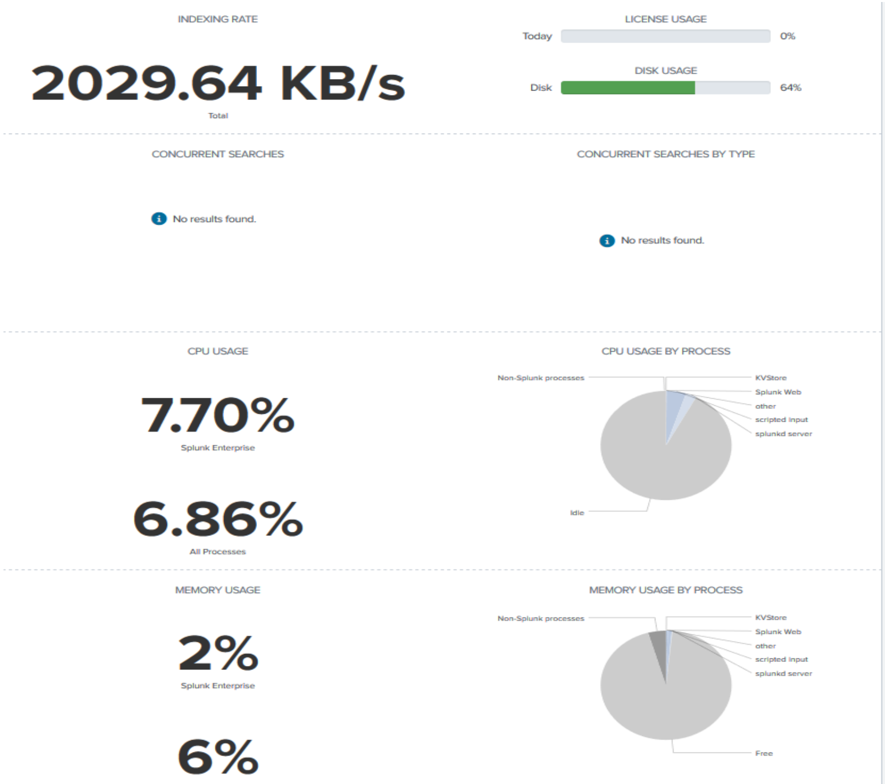
Figure14: hot, warm, cold bucket configuration

Volume: will indicate the label of the volume
Path: is the mount location
homepath: is the location of the hot and war bucket
coldpath: is the location of the cold bucket

In the LAB, Flash 9150 system running with the XFS file system are configured for the hot/warm buckets and IBM Elastic Storage System is configured for cold and frozen buckets using NFS protocol.

4. Monitoring the Splunk environment from the Splunk console

- **Indexing rate, CPU, Memory utilization from the Splunk Web Console**



INDEXING RATE

2029.64 KB/s

Total

LICENSE USAGE

Today 0%

DISK USAGE

Disk 64%

CONCURRENT SEARCHES

No results found.

CONCURRENT SEARCHES BY TYPE

No results found.

CPU USAGE

7.70%

Splunk Enterprise

6.86%

All Processes

CPU USAGE BY PROCESS

MEMORY USAGE

2%

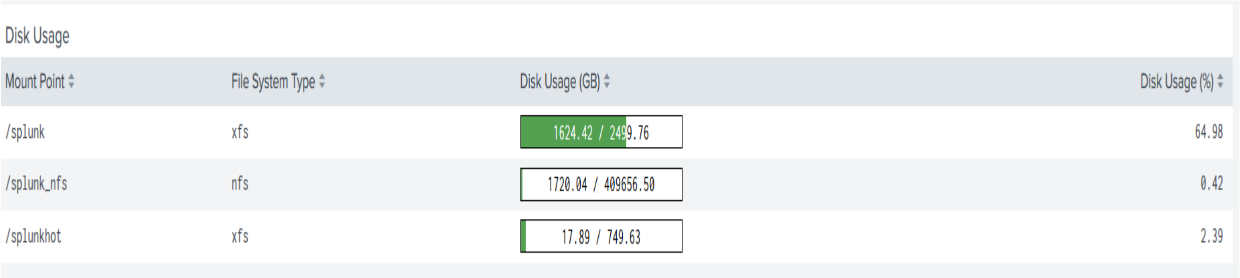
Splunk Enterprise

6%

MEMORY USAGE BY PROCESS

Figure15: Index rate and Compute resource monitoring using Splunk console

- **Disk Utilization.**

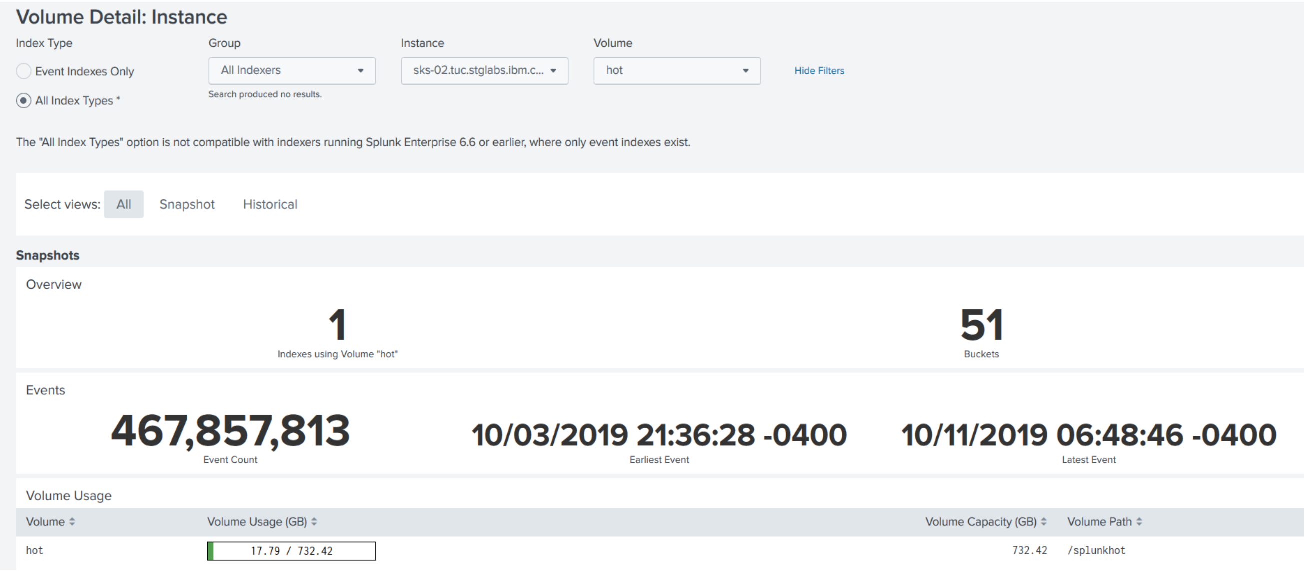


Disk Usage

Mount Point	File System Type	Disk Usage (GB)	Disk Usage (%)
/splunk	xfs	1624.42 / 2499.76	64.98
/splunk_nfs	nfs	1720.04 / 409656.50	0.42
/splunkhot	xfs	17.89 / 749.63	2.39

Figure16: Disk usage monitoring

- **Hot/warm volume details**



Volume Detail: Instance

Index Type
○ Event Indexes Only
● All Index Types *

Group: All Indexers
Search produced no results.

Instance: sks-02.tuc.stglabs.ibm.c...

Volume: hot

The "All Index Types" option is not compatible with indexers running Splunk Enterprise 6.6 or earlier, where only event indexes exist.

Select views: All Snapshot Historical

Snapshots

Overview

1
Indexes using Volume "hot"

51
Buckets

Events

467,857,813
Event Count

10/03/2019 21:36:28 -0400
Earliest Event

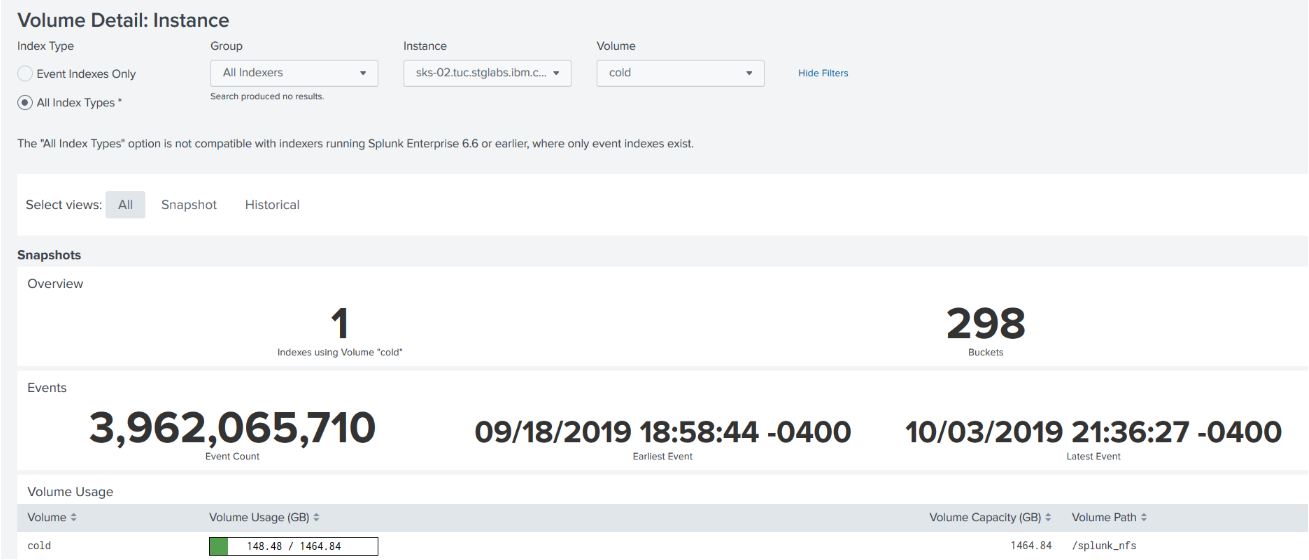
10/11/2019 06:48:46 -0400
Latest Event

Volume Usage

Volume	Volume Usage (GB)	Volume Capacity (GB)	Volume Path
hot	17.79 / 732.42	732.42	/splunkhot

Figure17: Hot/Warm Volume details

- **Cold volume details**



Volume Detail: Instance

Index Type

○ Event Indexes Only

● All Index Types *

Group: All Indexers

Search produced no results.

Instance: sks-02.tuc.stglabs.ibm.c...

Volume: cold

Hide Filters

The "All Index Types" option is not compatible with indexers running Splunk Enterprise 6.6 or earlier, where only event indexes exist.

Select views: All Snapshot Historical

Snapshots

Overview

1 Indexes using Volume "cold"

298 Buckets

Events

3,962,065,710 Event Count

09/18/2019 18:58:44 -0400 Earliest Event

10/03/2019 21:36:27 -0400 Latest Event

Volume Usage

Volume ⇕	Volume Usage (GB) ⇕	Volume Capacity (GB) ⇕	Volume Path ⇕
cold	148.48 / 1464.84	1464.84	/splunk_nfs

Figure 18: Cold Volume details

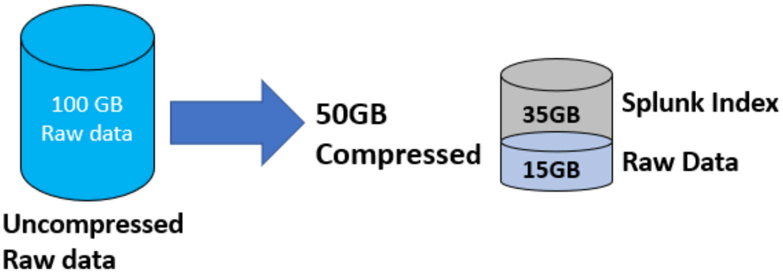
Storage Sizing Guide lines

This section provides some high- level storage configurations for the Splunk Enterprise Applications

I. **Hot/Warm Storage IO requirements per Indexer**
   - Around 800 IOPS for standard workloads
   - 1200 IOPS for Heavy workloads
   - Simultaneous sustained IO from multiple index nodes.
   - NFS/NAS is not recommended.
   - Size depends upon # of buckets X bucket size configured at the Splunk level.

II. **Cold/Frozen Storage IO requirements per Indexer**
   - up to 800 IOPS
   - File Storage is recommended
   - Size depends upon the retention period configured for the cold and frozen state.
   - Simultaneous sustained IO from multiple indexers

In general, Splunk will compress the incoming raw data by 50%.



Storage requirements in the Splunk Cluster environment:

Splunk offers Indexer clustering feature for high availability and prevents data in case of failures. Splunk replicates the data across indexers by keeping multiple copies of the incoming data. This requires additional storage on all indexer servers to store the replicated copies

Two factors – Replication factor and Search factor determines the data availability in the Splunk cluster deployments.

- **Replication factor**: This determines the number of copies of data the cluster maintains and therefore, the cluster's fundamental level of failure tolerance. Contains the raw data in the compressed format.
- **Search factor**: This determines the number of searchable copies of data the cluster maintains, and therefore how quickly the cluster can recover its searching capability after a peer node goes down. Indexed copy on the peer based on the replicated raw data.

Storage capacity is dependent on the replication and search factors configured at the cluster level.

Example 1:

a. Ingested data = 2TB; Retention Period =100 days; Replication factor=1; Search factor=1

Assuming 50 % compression; 15% raw data and 35% indexed data

Raw Data = Ingested Data x 15% x replication factor
= 2TB x 15% x 1
=300GB * 1
=300 GB

Indexed Data =Ingested Data x 35% x Search factor
=2TB x 35% x 1
= 700GB x 1
= 700 GB

Storage per day = Raw Data + Indexed Data
= 300GB + 500 GB
= 1TB

Total Storage = Storage per day * retention period
= 1 TB * 100
= 100 TB

Example 2:

Replication factor =3; Search factor =2; Daily Ingested data = 2TB; Retention period = 100 days
Assuming 50 % compression; 15% raw data and 35% indexed data

Raw Data = Ingested Data x 15% x replication factor
= 2TB x 15% x 3
=300GB * 3
=900 GB

Indexed Data =Ingested Data x 35% x Search factor
=2TB x 35% x 2
= 700GB x 2
= 1400 GB

Storage per day = Raw Data + Indexed Data
= 900 GB + 1400 GB
=2.3 TB

Total Storage = Storage per day * retention period
= 2.3 TB * 100
= 230 TB

Here is the sample sizing chart reference with various data ingest rates and indexers configured in the environment. Assuming Ingest rate of 100 GB/Day/Indexer; replication factor =3; Search Factor =2

Size	Indexers	Daily Ingest rate	Hot/warm	Cold	Frozen
		100	15 days	60 days	180 days
Small (S)	10	1 TB	17 TB	67 TB	79 TB
Medium (M)	50	5 TB	84 TB	337 TB	395 TB
Large (L)	500	50 TB	842 TB	3.4 PB	4 PB
Extra Large (XL)	1000	100 TB	1.7 PB	6.7 PB	7.9 PB
Double Extra Large (XXL)	1500	150 TB	2.5 PB	10 PB	12 PB

This is an illustration example only and actual sizing depends up on the ingest rates and retention periods configured in your environment. For detailed sizing guidelines, please refer the Splunk tool https://splunk-sizing.appspot.com/#adv=3&app=vmware&c=1&cdv=2&rl=10,01&rlv2=6&rlv3=6&st=v

Get More Information

This section lists the links to various manuals, red papers and publications for additional information

APPENDIX: sample configuration files

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) ®

Trademarking:

Refresh the IBM trademark list.

Run the **Toolkit** → **RXFM** → **Maintenance** → **Refresh-Toolkit-Rxfm-Rex-Scripts.rex** tool.

Trademark search and mark first use of a trademark:

Open the book file.

Run the **Toolkit** → **RXFM** → **Editor_tools** → **Trademark-Search.rex** tool.

The following terms are trademarks of other companies:

Other company, product, or service names may be trademarks or service marks of others.

Terms and conditions for product documentation

Permissions for the use of these publications are granted subject to the following terms and conditions.

Applicability

These terms and conditions are in addition to any terms of use for the IBM website.

Commercial use

You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not make derivative works of these publications, or reproduce, distribute or display these publications or any portion thereof outside your enterprise, without the express consent of IBM.

Rights

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any information, data, software or other intellectual property contained therein.

IBM reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by IBM, the above instructions are not being properly followed.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

IBM MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Privacy policy considerations

IBM Software products, including software as a service solutions, ("Software Offerings") may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user, or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering's use of cookies is set forth below.

This Software Offering does not use cookies or other technologies to collect personally identifiable information.

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, see IBM’s Privacy Policy at http://www.ibm.com/privacy and IBM’s Online Privacy Statement at http://www.ibm.com/privacy/details in the section entitled “Cookies, Web Beacons and Other Technologies,” and the “IBM Software Products and Software-as-a-Service Privacy Statement” at http://www.ibm.com/software/info/product-privacy.

June 2019

ISBN DocISBN

REDP-5549-00