Hedvig-Splunk Integration

Solution Brief

With the increasing growth in the volume of IT systems, as well as the continuous adoption of the cloud for deploying and seamlessly scaling these systems on demand, organizations are creating data at an accelerating pace. A significant contribution to this data growth is from sources outside the organizations, ranging from social media applications to real-time data from IoT devices.

Given the myriad of sources, the data generated does not conform to any predefined formats (unstructured), making it difficult to be stored in a traditional structures stores, such as databases, and interpreted differently under various contexts.

The inherent lack of structure makes it a challenge to search for and analyze useful business insights in the rush of unstructured data. Splunk creates value for organizations with an array of tools designed to efficiently process vast amounts of unstructured data. Splunk makes it easier and faster to collect data and then apply powerful indexing, search, and advanced analytics to provide real-time business operational intelligence.

While Splunk solves the challenge of helping organizations derive value out of unstructured data, it is necessary for it to be coupled with a storage system designed to seamlessly manage and store terabytes and petabytes of data in order to enable Splunk to do what it does best.

The Hedvig Distributed Storage Platform is designed to handle the high-performance and data-intensive workflows of Splunk across all stages of the Splunk data lifecycle. Hedvig powers in-software provisioning of file, block, and object storage with the flexibility to span private and public clouds, creating an elastic, hybrid cluster that can scale to thousands of nodes.

Data Management in Splunk

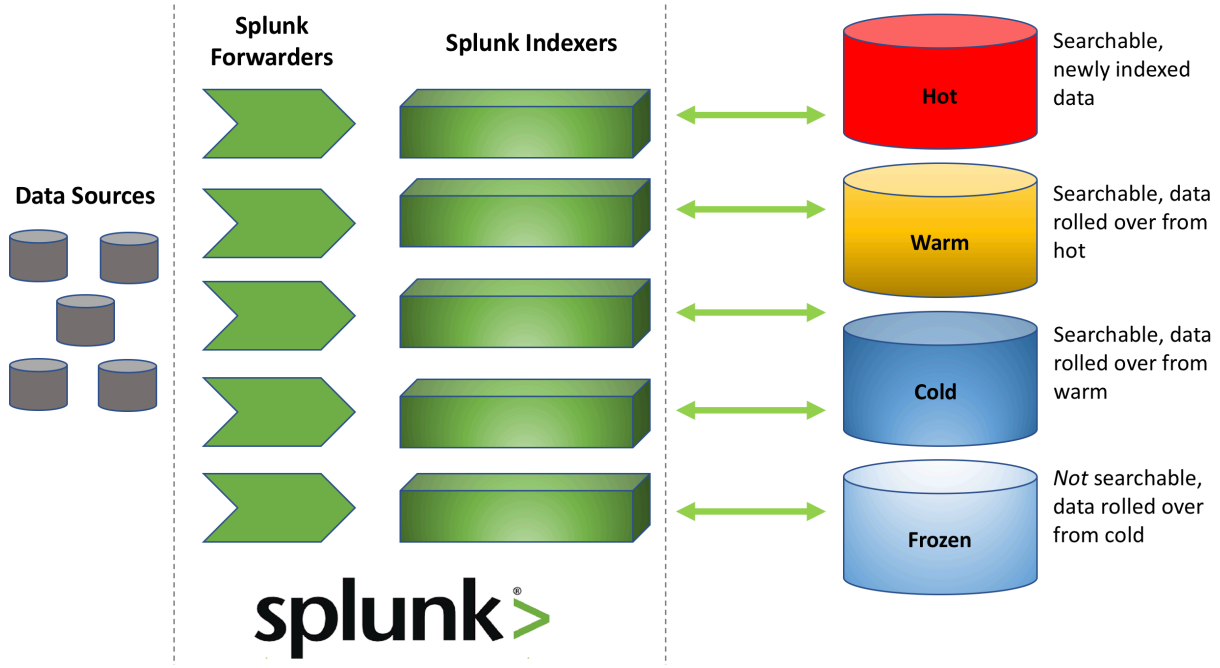
Splunk manages vast amounts of data through a big-data pipeline that consists of the following components:

- *Splunk Forwarders* consume and forward machine-generated data from practically any remote server, endpoint, or sensor.
- Data forwarded by Splunk Forwarders is collected and managed by *Splunk Indexers*. Splunk Indexers store raw machine-generated data in a compressed format, along with indexes pointing to the location of the data, thereby transforming machine data into searchable entities.
- *Splunk Search Head* allows users to query stored data by interfacing with indexers to gain access to the specific data they request.

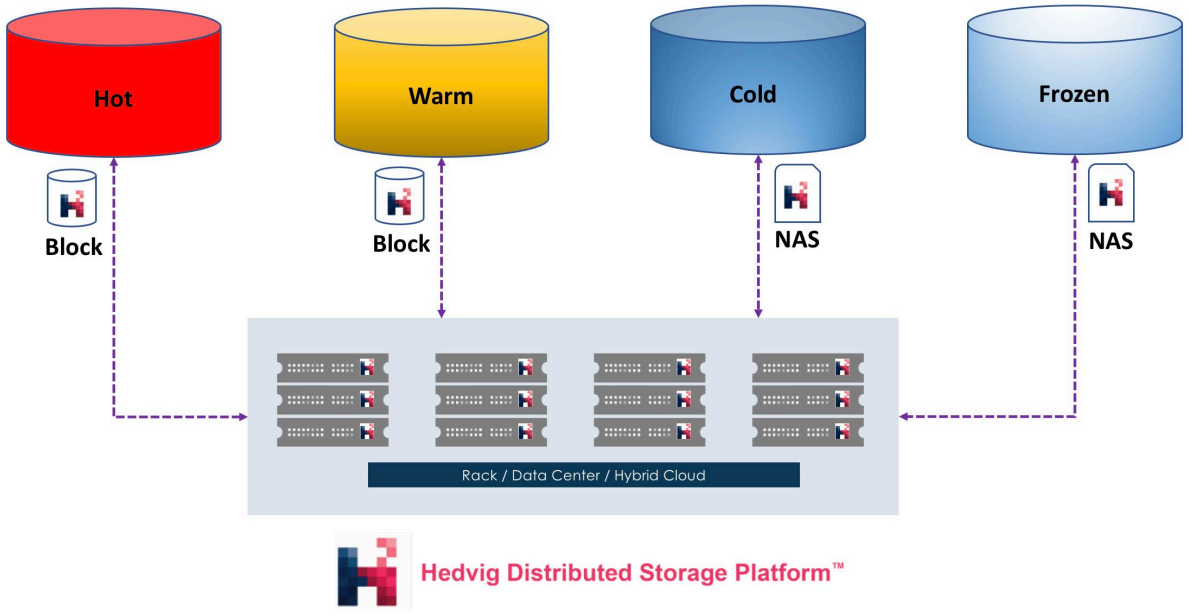
Splunk Data Storage and Hedvig

Splunk Enterprise stores indexed data in *buckets*. Each bucket can be defined as a collection of files within a directory. Each bucket moves through different stages as it ages. The different stages are defined as follows:

- **Hot** – The hot bucket stores newly indexed data. Once the hot bucket hits a size threshold or a time limit, it rolls over into a warm bucket, and a new hot bucket is created for new data. Data in the hot bucket is searchable.
- **Warm** – Warm buckets are searchable. Similar to hot buckets, when a size threshold or a time limit is hit, the warm bucket rolls over into a cold bucket.
- **Cold** – Cold buckets are searchable. Cold buckets roll over into a frozen bucket depending upon the retention period.
- **Frozen** – Frozen buckets are deleted by the indexer unless they are configured to be archived.



As Splunk aggregates unstructured data from multiple sources, storage volumes can quickly grow to terabytes or even petabytes with billions of files. Given the characteristics of different bucket stages, a hybrid storage environment of high-performance block storage and a scale-out, network-attached storage addresses these challenges. A single unified storage system that can meet the performance and capacity requirements of Splunk can greatly simplify storage management and reduce storage administration costs.



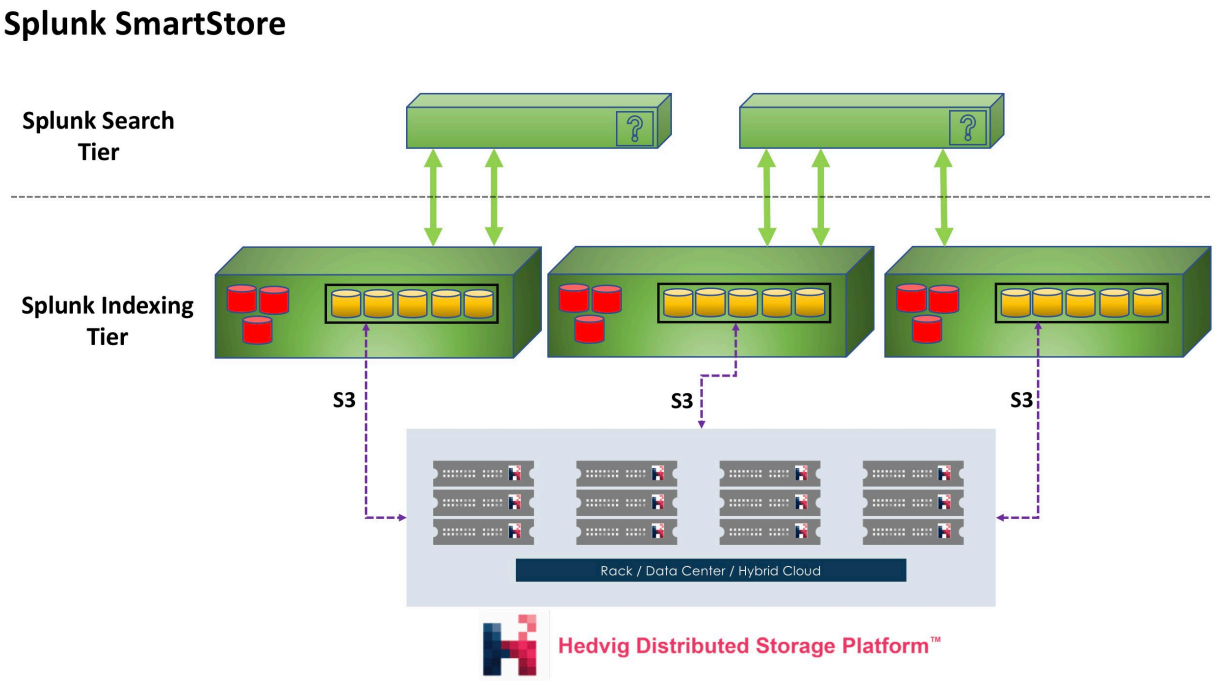
The Hedvig Distributed Storage Platform provides a modern solution with all the capabilities required to support Splunk's elastic and demanding workloads. Hedvig virtualizes and aggregates flash and spinning disk in a server cluster or cloud, presenting it as a single, elastic storage system. Organizations can provision any number of virtual disks fully customized to fit the needs of the applications.

Splunk hot and warm buckets are actively written and used during search operations. Splunk recommends a high-performance block storage tier for hot and warm tiers. Hedvig block virtual disks provisioned with the pin-to-flash option, coupled with deduplication and compression without performance penalties, provide a high-performance and low latency tier ideal for hot and warm buckets.

As the data ages, Splunk moves warm data to cold buckets. A low-cost storage tier with configurable retention policies, backed by a scale-out NAS share that can be expanded seamlessly to accommodate growing volumes, is ideal for cold and frozen tiers. Hedvig NFS virtual disks provide an ideal solution to store Splunk cold and frozen buckets. In addition, Hedvig scale-out NAS shares can also be leveraged for the Splunk data source tier that is read by Splunk Forwarders.

Splunk SmartStore

Splunk SmartStore is the latest evolution in Splunk's big data management paradigms. SmartStore dynamically places data in local storage, or in remote storage, or in both, depending on access patterns, data age, and data/index priority. SmartStore uses AWS S3 API to utilize any S3 API compliant object stores for the remote storage tier.



Hedvig's S3 compliant object storage is fully supported as a data storage tier for Splunk SmartStore. For existing Splunk deployments, leveraging Hedvig's object storage with SmartStore-enabled indexes can provide an economical solution by minimizing the use of expensive locally attached storage.

With SmartStore indexes, hot buckets are built in the indexer's local storage cache. As a bucket rolls from hot to warm, a copy of the bucket is uploaded to the Hedvig object store. The remote copy then becomes the master copy of the bucket. Eventually, the cache manager evicts the local bucket copy from the cache. When the indexer needs to search a warm bucket for which it does not have a local copy, the cache manager fetches a copy from the Hedvig object store and places it in the local cache.

Hedvig uses a combination of synchronous and asynchronous replication techniques to protect data across the cluster, spanning multiple disparate zones/regions/clouds to provide near-zero recovery point objectives (RPO) and recovery time objectives (RTO). This provides a significant boost over AWS S3/EBS native cross-region replication, in which objects are only eventually replicated, with replication times ranging between a few minutes to several hours, depending on the size and the number of objects.

Benefits of Hedvig

To summarize, the key benefits of Hedvig that enable enterprises to extract more value out of their business data with Splunk are:

- **A single storage platform with multi-protocol support** - The Hedvig Distributed Storage Platform eliminates the need for disparate primary and secondary storage solutions by providing native support for block, file, and object storage.

- **Advanced enterprise storage services** - The Hedvig Distributed Storage Platform provides a rich set of enterprise storage services, including deduplication, compression, encryption, snapshots, clones, replication, auto-tiering, multi-tenancy, and self-healing, to support production storage operations and enterprise SLAs.

- **Unmatched scale with performance optimization** - The Hedvig Distributed Storage Platform scales-out seamlessly with off-the-shelf commodity servers. Its superior metadata architecture and intelligent Client-Side Caching help to optimize performance for different workloads. A deployment can start with as few as three nodes and scale to thousands. Performance and capacity can be scaled up or down independently and linearly.

- **Multi-site disaster recovery** - The Hedvig Distributed Storage Platform inherently supports multi-site high availability, which removes the need for additional costly disaster recovery solutions. This empowers businesses to achieve native high availability for applications across geographically dispersed data centers.

- **Cloud-like simplicity with superior economics** - The Hedvig user interface provides intuitive workflows to streamline and automate storage provisioning. Admins can monitor and even provision storage assets from any device, including mobile devices, via a native HTML5 interface that does not require Flash or Java. This brings the provisioning simplicity of public clouds, such as AWS, to any data center.