

HPE Ezmeral Data Fabric Software



Contents

Introduction.....	3
HPE Ezmeral Data Fabric Software.....	3
Benefits of HPE Ezmeral Data Fabric Software.....	3
HPE Ezmeral Data Fabric Software components.....	7
File system.....	7
Object storage.....	7
Event streams.....	8
NoSQL database.....	8
Data resiliency, redundancy, and location.....	8
Snapshots and mirroring.....	8
HPE Ezmeral Data Fabric Software security.....	9
Conclusion.....	10



Introduction

Analytics has become the core component to every digital transformation but as data and the number of sources and formats continue to increase, organizations are struggling with some key challenges:

- A high level of complexity that comes from accessing, combining, security and managing data across different sources, formats, and providers
- This complexity results in data analytic teams spending most of their time with administrative tasks, such as discovering, copying, and normalizing data, rather than delivering insights to user expectations

Your business needs a new approach. One that delivers enterprise-wide visibility and direct access to hybrid data without having to move and then reformat the data before analysis begins.

HPE Ezmeral Data Fabric Software

HPE Ezmeral Data Fabric allows customers to conquer these hybrid challenges by replacing multiple point products with a single solution optimized for hybrid data analytics. It empowers users with corporate-wide visibility and direct access to data distributed across on-premises, in multiple clouds, and edge locations.

Designed to be target agnostic, this solution can be deployed on-premises, co-location, Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform™ (GCP™), HPE GreenLake, and edge locations. In essence, the solution allows users to see everything and access it from anywhere.

Benefits of HPE Ezmeral Data Fabric Software

Enterprise-wide data access

Direct access to global data is enabled by a native data plane that spans across all hybrid locations to combine files, objects, table data, and streaming data into a single trusted data source. The result is at-a-glance visibility and direct access to data from a single management view. Once set up, automated policies are created to consistently apply security, data sovereignty, and management configurations across all environments being managed by HPE Ezmeral Data Fabric. Setup, configuration, geo fencing, and management can also be established through this same management view.

Built into the data plane are two components, a global namespace and security management system, which work in tandem to ensure only authorized users can access hybrid data sets.

Global namespace

The global namespace combines multiple data types into a logical structure where servers, share, private and public cloud instances, and edge locations can access data through a single access point known as a namespace. As shown in Figure 1, the native security system works in tandem with the namespace to ensure only authorized users can view and access data regardless of location. This means that a user and the desired data set can be thousands of miles apart, but access is as simple as clicking an entry in the user's directory structure.

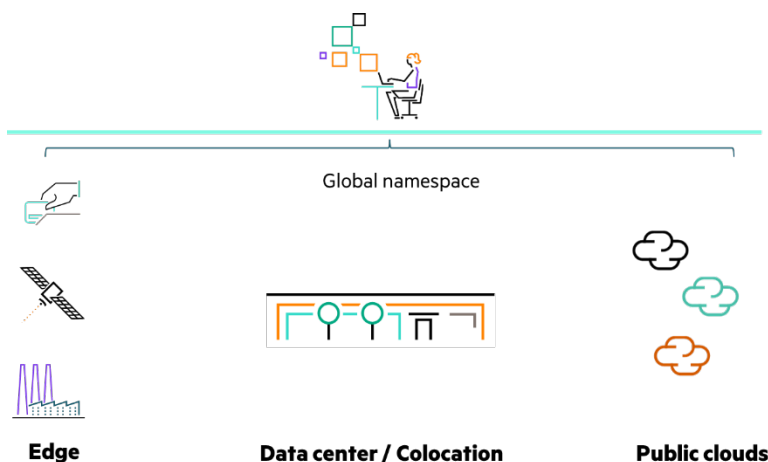


Figure 1. The global namespace delivers at-a-glance visibility and direct access across multiple locations



Once a namespace is established, data fabrics can be created using the built-in installer, shown in Figure 2. Usually, the criteria for creating a new data fabric are based on a firewall or network boundary. Hundreds of data fabrics can reside in a single namespace and a single data fabric can be configured to span across multiple namespaces.

The data plane stitches the multiple data fabrics together allowing direct access to the data as well as data sharing and team collaboration. Let's look at a real-world example.

Figure 2. Creating a new data fabric is a simple point and click process

Figure 3 represents a retail customer with over 200 sites spread across different locations. Data fabrics have been created for edge, cloud, and a centralized data center. Once trust is established, communication and data sharing can occur both within the fabric as well as across multiple data fabrics.

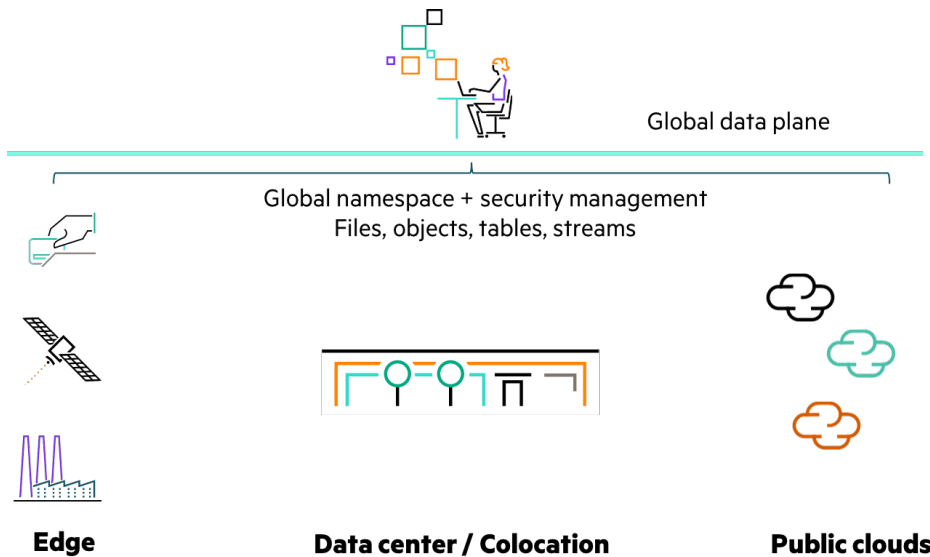


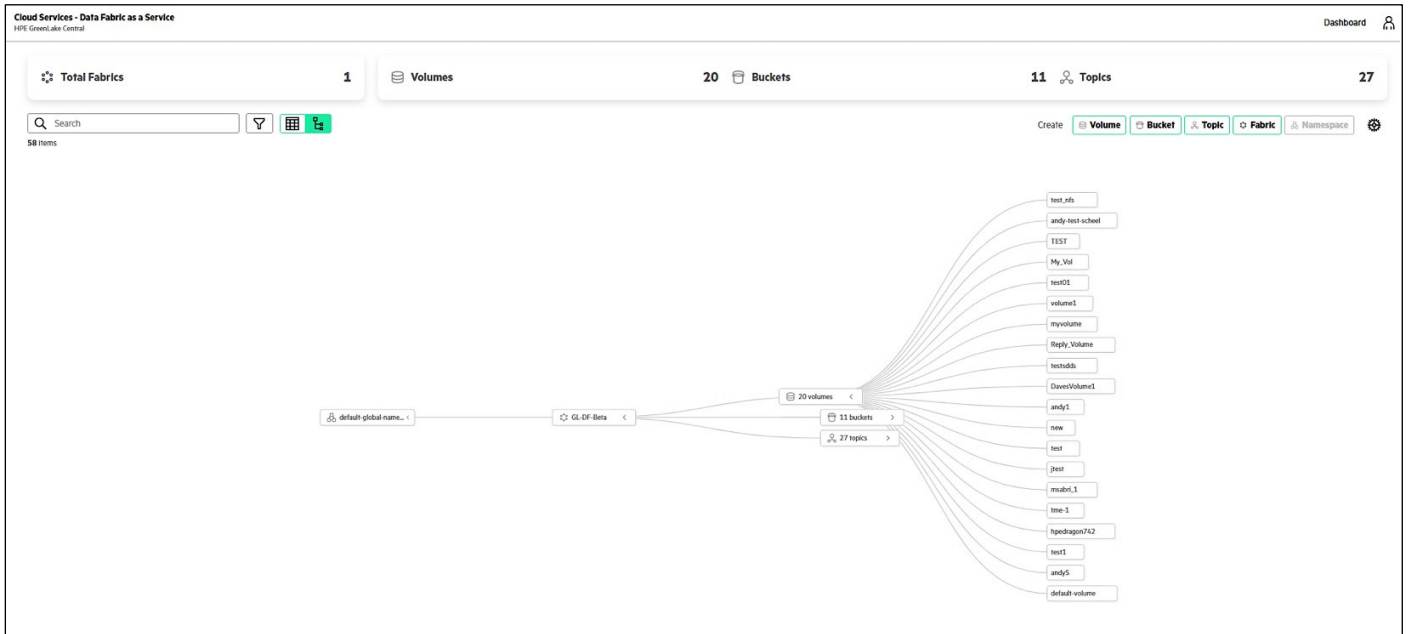
Figure 3. Retail example showing how data plane stitches together multiple data fabrics

This is useful if one store identifies a new fraud technique in the western United States and they want to share the insights, analysis, or video examples with other stores in their geography. A second use case would be replicating insights from real-time analysis into a centralized data center where it will be combined with data from other stores to identify new customer behavior patterns or buying trends.



Multiformat support

HPE Ezmeral Data Fabric Software As shown in Figure 4, clicking on the volume, bucket, or topic tab expands all the files under each specific category.



Note: this screenshot is from the SaaS UI.

Figure 4. Expansion of the volume category shows all the types of volumes configured on this data fabric

The top volume has a file written in NFS. A colleague wants to use an object-based application to read this data but neither the user nor the application has access rights. By clicking on the gear on the top far corner of the user interface, a listing of all authorized users is provided as shown in Figure 4. If the colleague's name is not in that listing, clicking on the gear again and choosing "invite user" solves the problem. Behind the scenes automation and orchestration establishes security and access controls and then informs the user they are now a member of the txt.nfs volume allowing them to read the data using their preferred protocol.

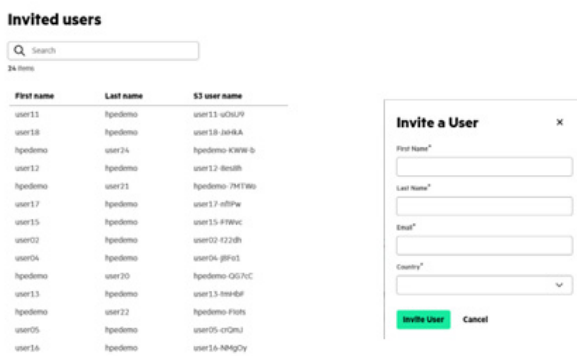


Figure 5. Point and click simplicity shows you have access to a volume and how to invite a new member

Simplified data management

Regardless of the user location, HPE Ezmeral Data Fabric simplifies data management and reduces the overhead of multiple analytic point solutions as follows:

- Increases data quality by aggregating multiple data types into a single source free of duplicates
- Intelligent mirroring reduces costs by lowering the retention of duplicate copies of data sets stored across multiple locations



- Accelerates time to Insights by writing files, objects, streams, and table data in their native format, then allows multiple applications to read in a different format; for example, write In S3 and read In Network File System (NFS).
- Helps ensure conformity with advanced features for geo-fencing, end-to-end encryption, and compliance with Industry and government regulations.

Built for hybrid data deployments

The edge and the ability to process data in real-time at the point of creation is growing as organizations realize the advantage of anticipating customer needs quickly to operate their business with greater efficiency and less waste. Designed for hybrid data deployments, HPE Ezmeral Data Fabric Software reduces costs by replacing multiple products with a single solution that secures, manages, and processes hybrid data that:

- Reduces time spent on data discovery, copying data across multiple locations, and Integration and normalizing of hybrid data
- Delivers a cost-efficient method with built-in mirroring for copying data sets and synchronizing across multiple locations reducing costs, latency, and performance issues
- From a single user interface setup and configure policies to consistently apply data access, lineage, security, and governance policies across different services and providers

Multi-API data access

By providing open, multi-API data access, the HPE Ezmeral Data Fabric Software enables legacy applications to work directly with modern analytics and AI applications while safely sharing the same data on the same system.

HPE Ezmeral Data Fabric Software allows the use of the full range of familiar Linux® commands. Apache Spark or Apache Hadoop programs can run unchanged. Kubernetes-based applications are also able to directly access and operate on data stored in the HPE Ezmeral Data Fabric Software using the Container Storage Interface (CSI) driver.

HPE Ezmeral Data Fabric Software supports a range of open data access protocols including POSIX-based APIs, Big Data APIs such as HDFS, and cloud-based APIs including Amazon Simple Storage Service (AWS S3). This data access flexibility frees developers and analysts to use the tool or language of their choice (Java, Python, Apache Spark, Apache Hive, SQL). Data scientists can use AI and machine learning tools to directly access data in the HPE Ezmeral Data Fabric Software without having to copy data to a special-purpose system. Universal data access by different applications is illustrated in Figure 6. The applications shown in Figure 6 would use the same pathname to access data despite using different APIs (See Figure 6).

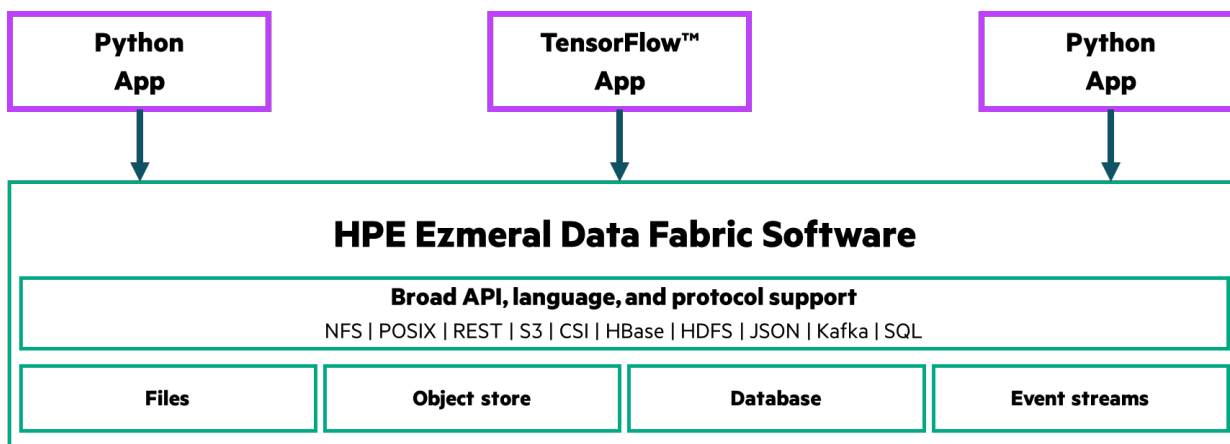


Figure 6. Direct access to data stored in HPE Ezmeral Data Fabric Software by applications using different APIs built with different tools

Global namespace

With HPE Ezmeral Data Fabric, all its files, objects, databases, and streams can be in a single global namespace even if they reside on separate clusters, in different locations. With a global namespace, an application can access data on a local HPE Ezmeral Data Fabric Software cluster or remotely access data stored on an HPE Ezmeral Data Fabric Software cluster in another location, including edge, on-premises, cloud, multicloud, and hybrid cloud deployments (Figure 7).



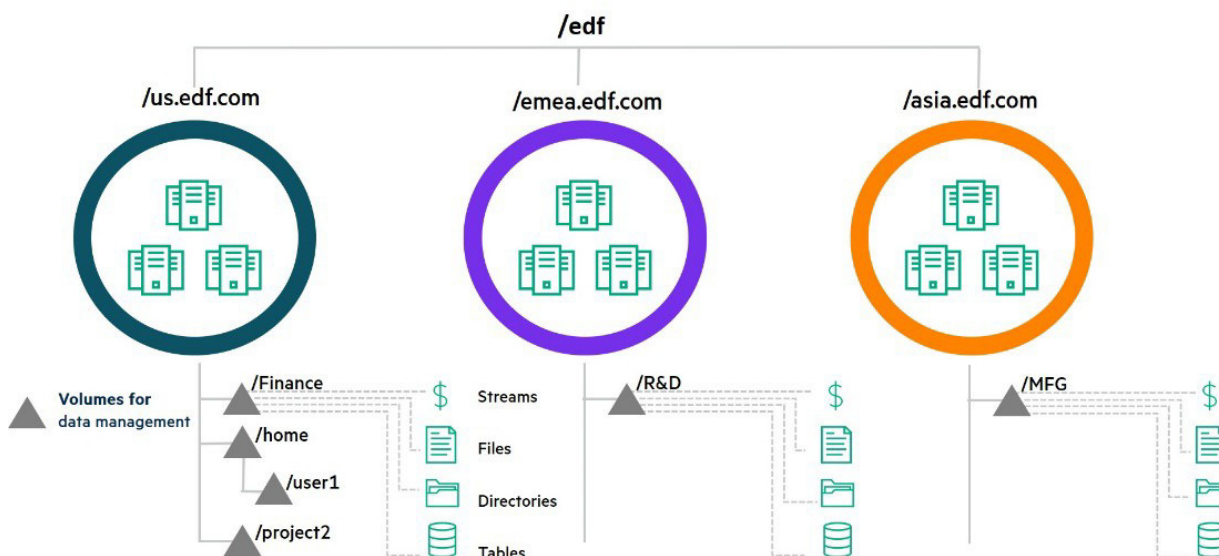


Figure 7. Local and remote data access enabled by the global namespace of HPE Ezmeral Data Fabric

With a global namespace, developers, analysts, and data scientists can reference data on remote clusters as if it were local. This easy access simplifies applications and makes it easier for multiple applications to work together on the same data. It also greatly simplifies the conceptual design of large systems.

Additionally, the global namespace capability establishes a separation of responsibilities, freeing those who build programs and models to focus on how their applications use data rather than having to be concerned with data logistics. This separation also allows system administrators to focus on the administration of the platform and less on application-level concerns.

Together, universal data access through multiple open APIs and an edge-to-cloud global namespace are key to how HPE Ezmeral Data Fabric Software serves as a unifying data infrastructure. Benefits include flexibility in tool and language choice, the ability to run legacy applications and modern applications on the same underlying data layer, and more efficient data management.

HPE Ezmeral Data Fabric Software components

As mentioned earlier, the HPE Ezmeral Data Fabric Software is more than just a file system, it can also store objects, and streams. This section explains in more detail the file store and the other components.

File system

HPE Ezmeral Data Fabric Software file system is a distributed file system for data storage, data management, and data protection. It supports mounting and cluster access through NFS and FUSE-based clients.

The file system is the only cloud-scale datastore that enables you to build a fabric of exabyte scale. It supports trillions of files, hundreds of thousands of client nodes, and runs on edge clusters, on-prem data centers, and the public cloud.

Object storage

The biggest advantage of object storage is its ability to add value to primary data through metadata or labels, enabling companies to easily search large data volumes, view audit trails, set policies, and provide auditable records on who can see, open, or download the data. The HPE Ezmeral Data Fabric Software provides key advantages over other object-based solutions:

- Optimized storage and performance for analytics by optimizing all object sizes for both performance and storage efficiency in a persistent datastore.
- Multiprotocol object access using the native S3 API or standard interfaces such as NFS, HDFS, and CSI allows data science applications and teams to use existing data access mechanisms.
- Resiliency and reliability inherited from the HPE Ezmeral Data Fabric Software with globally synchronized edge-to-cloud access while orchestrating clusters and data together. A global namespace simplifies edge-to-cloud topologies by incorporating traditional file storage and object store data into a single namespace.



Event streams

Data from many sources (producers) are written to topics, and many topics—potentially thousands or more—are bundled together in a single HPE Ezmeral Data Fabric Software stream. Customer applications read messages from streams (rather than the event data being broadcast to consumers), and reading a message does not cause a deletion. For that reason, customers can be added even long after the events are written to topics, providing considerable flexibility for the developer because the producers and consumers of messages can be highly decoupled and thus developed and deployed independently.

Each stream—potentially with thousands or even millions of topics—is essentially equivalent to an entire Kafka cluster. An HPE Ezmeral Data Fabric Software cluster can support millions of streams if needed. Many streams can be replicated into a single stream, even between clusters. These capabilities make the HPE Ezmeral Data Fabric Software useful in industrial IoT use cases where data from many edge sources must be collected and moved to central data centers, either partially processed or as raw data. Event streams in HPE Ezmeral Data Fabric Software streams can be accessed through the Apache Kafka API.

NoSQL database

There are two kinds of databases built into the HPE Ezmeral Data Fabric:

- Binary-style database
- Document-style database

Both kinds are as easy to create as it is to create a file. As mentioned in the discussion of the global namespace, HPE Ezmeral Data Fabric Software databases can be stored in any directory, wherever one would store a file. HPE Ezmeral Data Fabric Software databases have pathnames such as files and the permissions use the same type of access control expressions (ACEs). This means that volumes can be used to manage files, databases, and directories consistently.

The integration of databases into the distributed file system of the HPE Ezmeral Data Fabric Software allows a wide range of applications to be built on a single cluster, without the requirement to install, manage, or scale separate clusters for databases (or streams). The files in an HPE Ezmeral Data Fabric Software can also be used as the underlying storage for other, more conventional, kinds of databases including relational or NoSQL databases.

The way those databases are implemented in the HPE Ezmeral Data Fabric Software provides extreme scalability with excellent performance and provides significant flexibility. New databases or indexes can be added, or the data can be restructured easily, even at scale. Modern businesses need to be able to move quickly and respond to changing situations or changing business goals, so the flexibility afforded by HPE Ezmeral Data Fabric Software databases is an important feature of using it.

Data resiliency, redundancy, and location

HPE Ezmeral Data Fabric Software is inherently resilient and redundant due to its unique architecture and the built-in capability of replicating data across nodes and disks. Data stored in an HPE Ezmeral Data Fabric Software cluster can reside in any major public cloud, as well as in a private cloud, on-premises data centers, or edge locations. Data in any of these locations can be accessed directly or set up to mirror data to multiple locations to optimize for local access performance. This capability of HPE Ezmeral Data Fabric Software is a powerful advantage in being able to architect a system that is suited to current needs while still allowing modifications as needs change. The capability to access data remotely and to place data in multiple locations also means that applications can run wherever is best, depending on the type of hardware that is most suitable for the type of workload.

Snapshots and mirroring

HPE Ezmeral Data Fabric Software volumes are the basis for making true point-in-time snapshots of data and mirroring. Snapshots and mirrors can be created, and their lifetime can be set either through an automated schedule or manually.

Mirroring

HPE Ezmeral Data Fabric Software volumes are the basis for mirroring. HPE Ezmeral Data Fabric Software provides built-in mirroring to set recovery-time objectives and to automatically mirror data for backup. Local or remote mirror volumes can be created to mirror data between clusters, data centers, or on-premises and public cloud infrastructures. Mirror volumes are read-only copies of a source volume.

When a mirror volume is created, the HPE Ezmeral Data Fabric Software file system creates a temporary snapshot of the source volume. The mirroring process reads content from the snapshot into the mirror volume. The source volume remains available for read and write operations during the mirroring process. The initial mirroring operation copies the entire source volume. Subsequent mirroring operations only update the differences between the source volume and the mirror volume.



Mirror volumes can be promoted to read-write volumes. The main use case for this feature is to support disaster recovery scenarios in which a read-only mirror needs to be promoted to a read-write volume so that it can become the primary volume for data storage. In addition, read-write volumes that were mirrored to other volumes can be made into mirrors (to establish a mirroring relationship in the other direction). Read-write volumes can also be converted back to read-only mirrors.

Snapshots

The main use of snapshots is for data protection against user or application errors. Snapshots provide an excellent way to handle data versioning, a capability of value to data science teams since data scientists often need to keep an exact version of the data used to train AI and machine learning models. Snapshots are quick and easy to make and are efficient in terms of resource usage.

Snapshots enable the rollback to a known good data set and recover data always in case of data corruption or accidental deletions, without the help of storage administrators. A snapshot is a read-only image of a volume that provides point-in-time recovery. Snapshots only store changes to the data present in the volume, and as a result, make extremely efficient use of the cluster's disk resources. Snapshots preserve access to historical data and protect the cluster from user and application errors. A snapshot can be created manually, or the process may be automated with a schedule. Snapshots are stored in the snapshots directory. Snapshots can always be viewed from this directory.

New write operations on a volume with a snapshot are redirected to preserve the original data. Snapshots only store the incremental changes in a volume's data from the time the snapshot was created. The storage used by a volume's snapshots does not count against the volume's quota.

HPE Ezmeral Data Fabric Software security

Securing enterprise data is critical. To make securing data in clusters easy, the HPE Ezmeral Data Fabric Software has a data protection scheme built directly into the platform that is enabled by default, simplifying the process of protecting critical data.

Since data must be shared between nodes on the cluster, data transmissions between nodes, and from the cluster to the client are vulnerable to interception. Networked computers are also vulnerable to attacks where an intruder successfully pretends to be another authorized user and then acts improperly as that user. Additionally, networked machines share the security vulnerabilities of a single node. The HPE Ezmeral Data Fabric Software supports the ability to apply for protection directly as data enters and exits the platform.

Secure by default

HPE Ezmeral Data Fabric, which also includes EEP components, is secure out of the box on all new installations, ensuring all network connections require authentication and all data in motion is protected with wire-level encryption. HPE Ezmeral Data Fabric Software provides the ability to apply security protection directly for data as it comes into and out of the platform without requiring an external security manager server or a particular security plug-in for each ecosystem component. The security semantics are applied automatically on data being retrieved or stored by any ecosystem component, application, or user.

Authorization

Restricting an authenticated user's capabilities on the system, HPE Ezmeral Data Fabric Software provides sophisticated authorization controls to ensure that users can perform only the activities for which they have permissions, such as data access, job submission, cluster administration, and more. These permissions can be granted by an administrator through the browser-based control system management and monitoring interface or by using the command-line utilities.

Encryption

Restricting an external party's ability to read data. Encryption is used to avoid exposure to breaches, such as packet sniffing and theft of storage devices. In a secure HPE Ezmeral Data Fabric Software cluster, data transmission between nodes and between a data-fabric cluster and ecosystem application is encrypted, preventing an attacker with access to that communication from gaining information about the contents of the transmission. Data encryption at rest can be enabled to prevent unauthorized users from accessing sensitive data and protect against data theft through sector-level disk access.

Data is protected by encrypting all data being transmitted over the wire and optionally encrypting all that is stored on the data-fabric platform. The data-fabric data encryption scheme is built directly into the platform and is enabled by default.



Policy-based security

Create security policies and apply them to data objects to simplify the management of security controls on data. Policy-based security is a feature that administrators can use to classify security controls into a manageable number of security policies instead of defining security controls on individual data objects. The security controls defined in a security policy identify which users are authorized to access and modify data objects, whether to audit data operations, and whether to protect data in motion with wire-level encryption. When you apply security policies on data objects, such as volumes, files, and tables, the HPE Ezmeral Data Fabric Software automatically enforces the security controls defined in the policies during data operations. In cases where data is not associated with a security policy, the system enforces the security controls directly defined on data objects.

Single sign-on (SSO)

SSO is a method of authentication that allows users to access multiple applications or systems with a single set of login credentials. SSO is a valuable tool for streamlining the authentication process and reducing the burden on users who need to access multiple systems throughout their day-to-day work. It also improves security by reducing the number of login credentials.

Data Fabric implements SSO using the open-source Keycloak identity and access management (IAM) solution in a global namespace to allow signing on only once. SSO is used to access the data fabric, but users authenticate only once with Keycloak rather than the individual data fabrics. Keycloak is used only for access to the data fabrics in a GNS, not for access to EEP and other components.

Conclusion

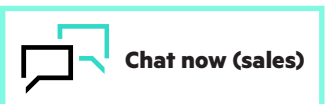
HPE Ezmeral Data Fabric Software is unique in the industry as a highly scalable and resilient data infrastructure that provides excellent performance for globally distributed systems including edge and hybrid cloud architectures. Global security and administrative policies ensure secure and consistent data sharing, making multi-tenancy practical and desirable.

The immediate advantages of using the HPE Ezmeral Data Fabric Software are straightforward, as described throughout this paper. HPE Ezmeral Data Fabric Software not only meets the criteria for an enterprise-wide unifying data infrastructure, but it also addresses the future.

Whether deployed as a software as a service (SaaS), or customer managed HPE Ezmeral Data Fabric Software reduces costs and frees data engineers, data scientists, analysts, and IT professionals from tedious infrastructure overhead, so they can focus on their tasks. The result is higher efficiency, faster development, reliability in production, and openness to innovation.

Learn more at
[HPE.com/datafabric](https://hpe.com/datafabric)

Explore **HPE GreenLake** 



© Copyright 2023 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Google Cloud Platform, GCP, and TensorFlow are registered trademarks of Google LLC. Linux is the registered trademark of Linus Torvalds in the U.S. and other countries. Azure and Microsoft are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Java is a registered trademark of Oracle and/or its affiliates. All third-party marks are property of their respective owners.