

Hitachi Data Lakehouse

Reference Architecture Guide

© 2025 Hitachi Vantara LLC. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including copying and recording, or stored in a database or retrieval system for commercial purposes without the express written permission of Hitachi, Ltd., Hitachi Vantara, Ltd., or Hitachi Vantara LLC (collectively "Hitachi"). Licensee may make copies of the Materials provided that any such copy is: (i) created as an essential step in utilization of the Software as licensed and is used in no other manner; or (ii) used for archival purposes. Licensee may not make any other copies of the Materials. "Materials" mean text, data, photographs, graphics, audio, video and documents.

Hitachi reserves the right to make changes to this Material at any time without notice and assumes no responsibility for its use. The Materials contain the most current information available at the time of publication.

Some of the features described in the Materials might not be currently available. Refer to the most recent product announcement for information about feature and product availability, or contact Hitachi Vantara LLC at https://support.hitachivantara.com/en_us/contact-us.html.

Notice: Hitachi products and services can be ordered only under the terms and conditions of the applicable Hitachi agreements. The use of Hitachi products is governed by the terms of your agreements with Hitachi Vantara LLC.

By using this software, you agree that you are responsible for:

1. Acquiring the relevant consents as may be required under local privacy laws or otherwise from authorized employees and other individuals; and
2. Verifying that your data continues to be held, retrieved, deleted, or otherwise processed in accordance with relevant laws.

Notice on Export Controls. The technical data and technology inherent in this Document may be subject to U.S. export control laws, including the U.S. Export Administration Act and its associated regulations, and may be subject to export or import regulations in other countries. Reader agrees to comply strictly with all such regulations and acknowledges that Reader has the responsibility to obtain licenses to export, re-export, or import the Document and any Compliant Products.

Hitachi and Lumada are trademarks or registered trademarks of Hitachi, Ltd., in the United States and other countries.

AIX, DB2, DS6000, DS8000, Enterprise Storage Server, eServer, FICON, FlashCopy, GDPS, HyperSwap, IBM, IntelliMagic, IntelliMagic Vision, OS/390, PowerHA, PowerPC, S/390, System z9, System z10, Tivoli, z/OS, z9, z10, z13, z14, z15, z16, z17, z/VM, and z/VSE are registered trademarks or trademarks of International Business Machines Corporation.

Active Directory, ActiveX, Bing, Excel, Hyper-V, Internet Explorer, the Internet Explorer logo, Microsoft, Microsoft Edge, the Microsoft corporate logo, the Microsoft Edge logo, MS-DOS, Outlook, PowerPoint, SharePoint, Silverlight, SmartScreen, SQL Server, Visual Basic, Visual C++, Visual Studio, Windows, the Windows logo, Windows Azure, Windows PowerShell, Windows Server, the Windows start button, and Windows Vista are registered trademarks or trademarks of Microsoft Corporation. Microsoft product screen shots are reprinted with permission from Microsoft Corporation.

All other trademarks, service marks, and company names in this document or website are properties of their respective owners.

The open source content used in Hitachi Vantara products may be found within the Product documentation or you may request a copy of such information (including source code and/or modifications to the extent the license for any open source requires Hitachi make it available) by sending an email to OSS_licensing@hitachivantara.com.

Feedback

Hitachi Vantara welcomes your feedback. Please share your thoughts by sending an email message to Docs-Feedback@hitachivantara.com. To assist the routing of this message, use the paper number in the subject and the title of this white paper in the text.

Thank you!

Revision history

Changes	Date
Initial release	December 2025

Reference Architecture Guide

Purpose

This reference architecture provides a practical blueprint for building the Hitachi Data Lakehouse-powered VSP Storage Platform One (VSP One) – a modern, unified, hybrid-cloud storage solution for data management. This guide is written for solution architects, platform and infrastructure engineers, and IT decision makers who want a unified, secure way to analyze data across databases, files, and object stores with minimal data movement.

At a high level, the guide explains how Hitachi Vantara's VSP One Block, VSP One File, and VSP One Object work together with Hitachi Advanced Database (HADB), Zetaris, and Pentaho Platform to deliver a comprehensive Data Lakehouse solution.

Key solution benefits

- Easy and faster access to data: Removes the barrier to connecting data sources with no-code/low-code source integration
- Accelerated insight: Removes the burden of long tail data preparation by enabling faster data integrations
- Optimized cost: Empowers customers to choose the right storage tier and tooling for each workload with flexible architectural options.

By combining federated queries, a shared semantic layer, strong governance, and storage optimization, the Hitachi Data Lakehouse accelerates data analytics on current and historical data while maintaining clear operational boundaries and lowering complexities.

Guide benefits

- Architecture and deployment model: How compute, network and storage components fit together across Block, File and Object storage.
- Bill of materials: An inventory of the hardware and software components you need – and why each matter.
- Data flow and orchestration: How queries run across Block, File, and Object, and how governance is enforced.
- Validation baseline: Representative use cases and test scenarios you can reproduce in the lab.
- Operational guidance: Day-to-day operations, guardrails and considerations for scale-out, so you can move from pilot validation to broader deployment with confidence.

Scope and assumptions

This document focuses on design and validation. It is not an installation manual or a formal performance benchmark. Version specifics and sizing should be confirmed against current product documentation and support matrices during planning.

Outcome

Following this architecture, organizations can unify data access, reduce operational complexity and unlock actionable insights - with minimal disruption to existing infrastructure and a clear path from proof-of-concept to production.

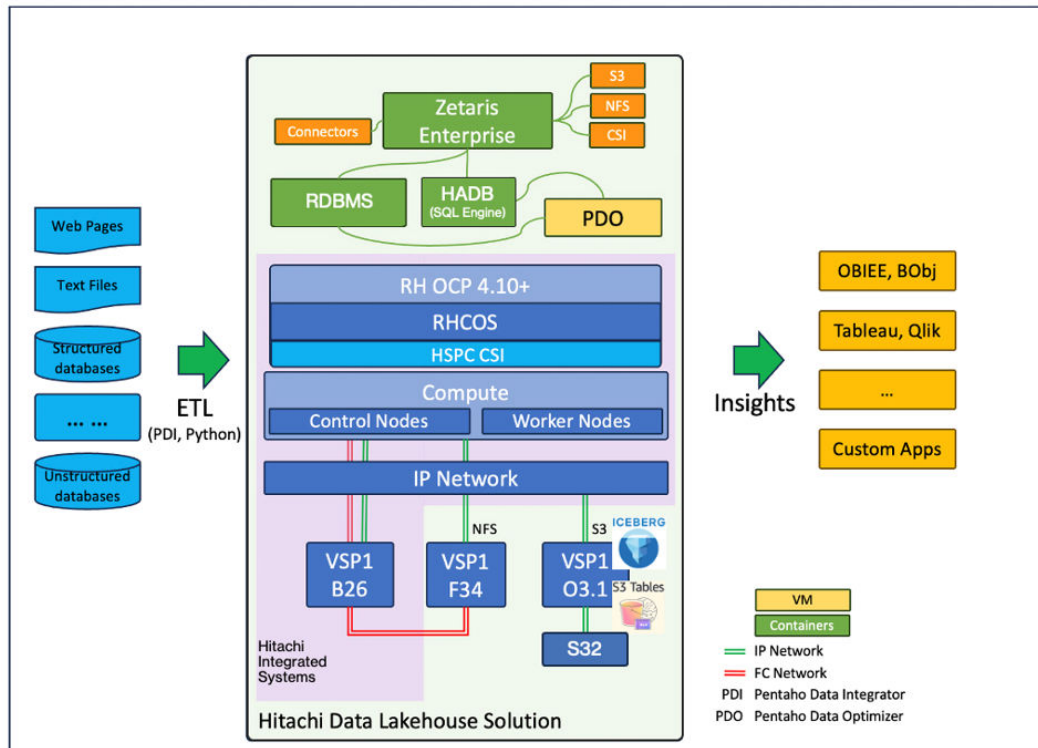
Solution overview

The Hitachi Data Lakehouse provides a modern, secure way to analyze data across databases, files, and object storage – with minimal data movement. It is designed to eliminate data silos and deliver real time insights while keeping strong governance in place.

At its core, the solution brings together three Hitachi Vantara storage platforms:

- VSP One Block for high performance databases and other stateful services
- VSP One File for shared files, logs, and collaborative assets
- VSP One Object (with S Series nodes where applicable) for long-term and unstructured data, including Parquet datasets for analysis.

At a high level, the following illustration captures the essence of the solution. Hitachi Integrated Systems, which is comprised of an orchestration layer, compute, network, and Block storage, is extended with file and object storage to offer a comprehensive foundation for Data Lakehouse.



Zetaris Enterprise

Zetaris Enterprise acts as the federated query engine and unified semantic layer, so users can query across Block, File, and Object through a single, governed interface. For large historical or compliance datasets stored as Parquet on Object, Hitachi Advanced Database (HADB) provides in place SQL (via S3 foreign tables), avoiding the need to reload data into a separate database.

Hitachi Advanced Database

HADB enables high-performance, in-place SQL analytics on large historical datasets stored in VSP One Object, allowing users to query external data directly from object storage without the need for ingestion or duplication. This capability streamlines archival data access and supports unified analytics across both current and historical data sources within the lakehouse architecture.

Pentaho

Pentaho Data Optimizer (PDO) automates data lifecycle management and intelligent tiering by classifying, archiving, and migrating low-value or inactive data from high-cost storage to cost-efficient object stores such as VSP One Object. This helps organizations reduce storage costs, improve compliance, and optimize data accessibility across hybrid and cloud environments within the Hitachi Data Lakehouse architecture.

The solution runs on Red Hat OpenShift (OCP), with most components deployed as containers. Storage integration is standardized on Hitachi Block and NFS CSI drivers for Block and File, and S3 for Object. For host-to-Block connectivity, the design prioritizes Fibre Channel for use with the Hitachi Block Storage Plug-in for Containers.

Key capabilities include:

- Federated querying & semantic governance — Unified, governed access to distributed data; works with BI tools such as Tableau and Qlik.
- Policy driven archival capabilities – to automatically move inactive data to cost-efficient object storage via Pentaho Data Optimizer.
- In place analytics on Object — Query Parquet on Object through HADB without re-ingest and combine “hot” and “cold” data in one view from Zetaris.
- Collaboration at scale — Shared file services on VSP One File support multi team notebooks, logs, and artifacts on OCP.
- Modular deployment — Single rack to multi rack configurations with flexible compute and storage options.
- Security & governance — Centralized policies, minimized data copies, and features such as object lock/legal hold on VSP One Object.



Note: Pentaho Data Optimizer is optional. It can be introduced later without changing the overall architecture or access patterns.

This solution helps organizations unify their data landscape, simplify operations, and speed up time to insight—while maintaining enterprise grade security and control on OpenShift and VSP One storage.

Business overview

Modern enterprises are drowning in data – growing fast, arriving in many formats, and spread across multiple systems. Traditional solutions like data warehouses for structured data and data lakes for everything else address parts of this challenge, but they struggle to deliver one, governed view of the truth in near real time. Instead, they often create more complexity – more copies, more pipelines and more overhead to manage.

The Hitachi Data Lakehouse addresses this by letting organizations analyze data with minimal data movement. Zetaris provides a unified semantic layer and federated SQL so teams can query databases, files, and object storage without bulk data movement. For large datasets stored as Parquet in object or file storage, Hitachi Advanced Database (HADB) enables in place SQL, avoiding re ingest and keeping costs predictable. Together, this reduces data sprawl and shadow IT, improves governance, and shortens the path from question to answer.

Underneath, the solution pairs VSP One Block (for high performance databases and stateful services), VSP One File (for shared files, logs, and team collaboration), and VSP One Object (for long term and unstructured data). Deployed on Red Hat OpenShift, most components run as containerized services with clear operational boundaries, so platform teams can manage and secure them consistently. Optional tools – such as Pentaho Data Optimizer for policy based tiering – can be added later without changing how users access data today.

This reference architecture demonstrates tangible business value by helping organizations:

- Unify access to structured and unstructured data in one governed view
- Speed time to insight by running queries at source freshness (not after nightly ETL)
- Reduce operational complexity and cost by minimizing pipelines and copies
- Strengthen compliance with centralized controls and fewer uncontrolled datasets
- Adopt a modular path from lab validation to broader rollout as needs grow.

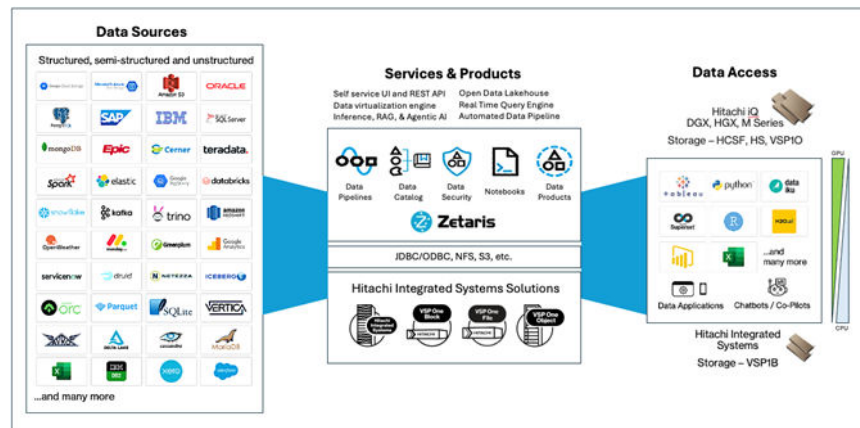
In short, Hitachi and Zetaris provide a secure, practical way to bring all your data together—without moving it—so teams can make faster, more confident decisions on a platform that's ready to evolve with the business.

Key components and technologies

The Hitachi Data Lakehouse Solution integrates a modular, scalable, and secure set of hardware and software technologies to deliver unified data access, real-time analytics, and simplified governance across enterprise environments. This section outlines the essential components and their roles within the solution.

Zetaris Enterprise

Zetaris Enterprise is a networked data platform that delivers virtual data warehousing and data mesh capabilities by connecting directly to disparate data sources – databases, data lakes, APIs, file systems, and streams. Instead of relying on complex ETL pipelines and centralized data lakes, Zetaris enables federated query execution across hybrid and multi-cloud environments in real time, applying governance, security, and quality rules at the source. Its heterogeneous query optimizer decomposes SQL queries into intelligent sub-queries executed on each source system, with in-memory caching for performance, and unifies results into virtual views accessible to BI, AI/ML, and analytics tools. By virtualizing data access, Zetaris solves problems of data silos, costly duplication, and shadow IT while providing consistent governance, metadata lineage, and the ability to seamlessly blend historical and real-time, operational data for analytics and AI-driven applications.



Zetaris Enterprise Capabilities:

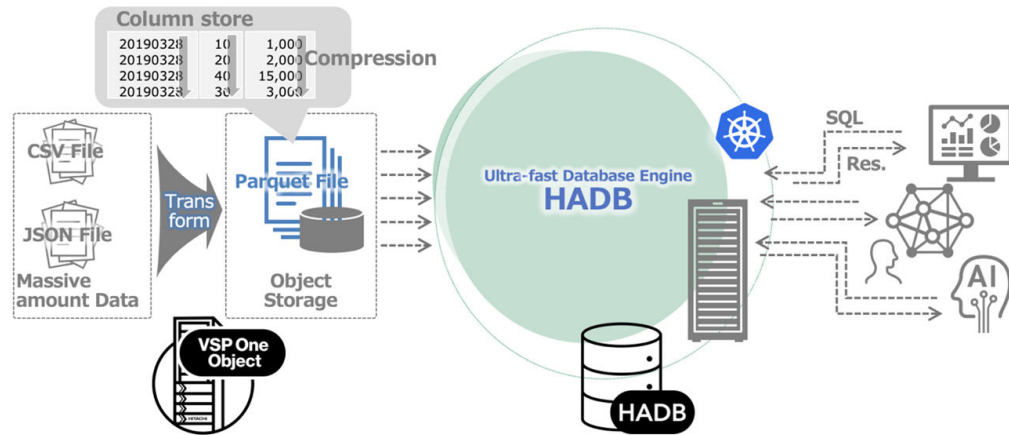
- Virtual Data Warehousing & Data Mesh – Federates queries across distributed data sources (databases, lakes, APIs, streams, files) without moving or duplicating data.
- Heterogeneous Query Optimizer – Breaks down SQL queries into intelligent sub-queries, executes them at source, and unifies results; applies dynamic optimization and caching for performance.
- Streaming + Historical Data Integration – Blends in-stream data with historical warehouse/lake data for real-time analytics and AI-driven applications.
- Data Catalog & Metadata Management – Captures table and column metadata, tracks lineage and statistics, and enables tagging, search, and governance.

To learn more about Zetaris Enterprise, reach out to your Hitachi account team or visit <https://www.zetaris.com>.

Hitachi Advanced Database

Hitachi Advanced Database (HADB) is a high-performance relational database optimized for large-scale analytics and near real-time data processing. With high-performance, multi-protocol storage support, HADB enables highly concurrent analytics across block storage (VSP One Block), file storage (VSP One File/NFS), and object storage (VSP One Object / S3-compatible). It employs an out-of-order parallel execution engine to maximize multi-core CPUs, memory, and SSD/flash, enabling fast joins, aggregations, recursive queries, and index searches without specialized hardware. HADB supports both row-store and column-store formats, allowing flexible optimization for transactional or analytical workloads.

A key capability for modern data lakehouse architectures is its support for S3 and file based foreign tables, which allow external data in CSV, Parquet, or JSON formats to be directly queried in place using SQL, without requiring ingestion or duplication. HADB supports Apache Iceberg for seamless integration with VSP One Object S3 Table Buckets. This enables many helpful Iceberg features such as ACID transaction enforcement over object based table data, time travel, and schema evolution. Partition-aware definitions further optimize performance by pruning S3 data directories during query execution. With high-speed parallel data imports, real-time query execution, and seamless integration across block, file, and object storage, HADB addresses the challenges of siloed storage, slow batch ETL, and data latency—delivering consistent, low-latency query performance across hybrid and cloud-native environments.



HADB Capabilities:

- High-Performance Query Engine – Out-of-order parallel execution for joins, aggregations, recursive queries, and index scans to maximize CPU and memory use.
- Hybrid Storage Model – Supports both row-store and column-store formats, enabling optimization for transactional lookups or analytical workloads.
- S3 Foreign Table and Iceberg Support – Directly query external data in S3 (CSV, Parquet, JSON) without ingestion; partition-aware design enables directory pruning for faster access.
- Real-Time Analytics – Supports continuous ingestion and query execution, making it suitable for scenarios like fraud detection, sensor monitoring, and streaming analytics.

To learn more about Hitachi Advanced Database, reach out to your Hitachi account team or visit <https://www.hitachi.com/products/it/software/prod/hadb/index.html>.

Pentaho Data Optimizer (PDO)

Pentaho Data Optimizer (PDO) is a data lifecycle management and intelligent tiering solution that integrates with the Pentaho Data Catalog to reduce redundant, obsolete, and trivial (ROT) data and optimize storage usage across hybrid, cloud, and on-premises environments. It uses rule-based classification and policy enforcement to identify low-value or inactive data, automatically tiering or archiving it from high-cost primary storage to cost-efficient object stores such as Hitachi Content Platform (HCP), VSP One Object, or S3-compatible targets. PDO supports tiering from NFS, SMB, SharePoint, OneDrive, and cloud sources, enabling transparent offloading without disrupting applications. It provides auditability, rehydration, and retention controls, ensuring data movements are reversible, compliant, and traceable. By aligning storage decisions with usage, business value, and compliance risk, PDO solves problems of storage cost inflation, data sprawl, and unmanaged risk exposure, while improving agility for analytics and AI initiatives.

Pentaho Data Optimizer (PDO) Capabilities:

- Intelligent Data Lifecycle Management – Automates classification, tiering, archiving, and purging of data based on business value, usage, and compliance rules.
- Rule-Based Governance: Enforces retention, access, and storage location policies across files, objects, and unstructured data.
- Multi-Protocol Source Integration: Supports NFS, SMB/CIFS, HDFS, SharePoint, OneDrive, AWS S3, Azure Blob, Hitachi Content Platform (HCP), and any S3-compatible storage.
- Auditability and Compliance: Provides full traceability and reversibility of data movements, with retention controls and audit logs aligned to PDC's governance framework.

For more information about Pentaho and Pentaho Data Optimizer, reach out to your Hitachi account team or visit <https://pentaho.com/>.

Red Hat OpenShift Container Platform (OCP)

Red Hat OpenShift Container Platform (OCP) is an enterprise-grade Kubernetes distribution that provides a secure, automated platform for deploying, managing, and scaling containerized applications across hybrid and multi-cloud environments. Built on Kubernetes and Red Hat Enterprise Linux CoreOS, OCP integrates container orchestration with features such as operator-driven lifecycle management, integrated CI/CD pipelines, service mesh, and monitoring/logging stacks. It enforces role-based access control (RBAC), security policies, and multitenancy isolation, addressing compliance and governance requirements critical for enterprise workloads. OpenShift simplifies day-2 operations through automated cluster scaling, self-healing, and rolling upgrades, reducing the operational complexity of managing large-scale distributed systems. By standardizing container infrastructure, OCP solves problems of inconsistent application deployment, operational overhead of bare Kubernetes, and lack of enterprise security and governance controls, while enabling consistent deployment of cloud-native applications – including data-intensive Lakehouse components – on-premises or across public clouds.

For more information about Red Hat OpenShift, visit <https://www.redhat.com/>.

Hitachi Advanced Server HA815 G3

Hitachi Advanced Server HA815 G3 is a 1U, dual socket platform based on 4th Generation AMD EPYC™ processors (up to 96 cores per CPU) with 24 DDR5 RDIMM slots (up to 6 TB total memory), 128 lanes of PCIe 5.0, and two OCP 3.0 slots for flexible I/O expansion. The chassis supports up to ten 2.5 inch hot plug NVMe, SAS, or SATA drives plus a dedicated boot device, with redundant hot plug power supplies and seven performance fans for enterprise class reliability, availability, and serviceability (RAS). System management is provided by the integrated iLO 6 ASIC with a dedicated 1 GbE management port for remote monitoring and provisioning. This balance of compute density, I/O bandwidth, and serviceability makes the HA815 G3 well suited for dense OpenShift clusters hosting data lakehouse services.



Hitachi Advanced Server HA815 G3 Capabilities:

- High-Core Density Compute – Supports one or two 4th Gen AMD EPYC™ processors with up to 96 cores per CPU, maximizing parallel processing for analytics, AI/ML, and virtualization workloads.
- Large Memory Capacity – Up to 6TB DDR5 memory across 24 DIMM slots (12 channels per processor), delivering high bandwidth for memory-intensive applications.
- Next-Gen I/O Bandwidth – Provides 128 PCIe Gen 5.0 lanes and 64 I/O lanes with CXL 1.1+ support, enabling ultra-fast interconnects for GPUs, NVMe storage, and networking.
- Flexible Storage Options – Supports up to 10 hot-plug NVMe, SATA, or SAS drives (max capacity 153.6TB), plus optional dual M.2 boot drives for OS and application resiliency.

To learn more about Hitachi servers, reach out to your Hitachi account team or visit <https://www.hitachivantara.com/>.

VSP One Block

Hitachi Virtual Storage Platform One Block B26 is an enterprise-class, midrange block storage system designed to consolidate mission-critical workloads on a single, scalable NVMe all-flash platform. Powered by symmetric active-active controllers and supporting up to 9.1PB effective capacity with NVMe SSDs, the B26 delivers <1ms latency and up to 256GB/s Fibre Channel bandwidth to handle both legacy databases and modern applications. It integrates patented adaptive data reduction with hardware-accelerated compression to maximize usable capacity without impacting performance. Enterprise-grade features include 100% data availability guarantees, global-active device clustering for active-active replication, immutable snapshots for ransomware protection, and dynamic drive protection to maintain resilience under drive failures. By unifying diverse workloads and eliminating siloed storage systems, the VSP One Block B26 solves problems of infrastructure sprawl, unpredictable performance, and data protection gaps, while enabling sustainable operations with Dynamic Carbon Reduction (DCR) energy optimization and ESG-aligned guarantees.



VSP One Block Capabilities:

- NVMe All-Flash Architecture – Delivers high throughput and low latency with scalable NVMe SSDs.
- Scalability – Supports up to 9.1PB effective capacity and expansion with additional NVMe drive shelves.
- Symmetric Active-Active Controllers – Ensures continuous availability and balanced workload distribution.
- High-Speed Connectivity – Up to 256GB/s Fibre Channel bandwidth, with support for FC-NVMe (16/32/64Gb), iSCSI (10/25Gb), and TCP (100Gb) interfaces.

To learn more about VSP One Block, reach out to your Hitachi account team or visit <https://www.hitachivantara.com/>.

VSP One File

Hitachi Virtual Storage Platform One File 34 is an enterprise-grade, hybrid-cloud ready file storage platform that delivers scalable performance, strong data protection, and simplified management for unstructured data workloads. It supports up to four nodes per cluster with aggregate bandwidth of 238 GB/s and throughput of up to 7 GB/s read and 1.9 GB/s write per node, making it suitable for analytics, AI/ML pipelines, and high-throughput file services. The system offers multiprotocol support—including SMB, NFS, FTP, iSCSI, and S3 tiering—and provides namespace scalability up to 4,160 billion files, consolidating large-scale unstructured datasets under a single architecture. Enterprise-grade security features include immutable snapshots, ransomware detection, multi-factor authentication, RBAC, and CyberArk privileged access management. Managed through the VSP 360 AI-powered control plane and SVOS data plane, it ensures consistent data mobility and governance across edge, core, and cloud. By combining high bandwidth, scale-out clustering, multiprotocol support, and built-in cyber resilience, VSP One File 34 addresses the challenges of unstructured data growth, hybrid-cloud integration, and ransomware defense while ensuring always-on availability with Hitachi's 100% Data Availability Guarantee.



VSP One File Capabilities:

- Cluster Scalability – Supports up to 4 nodes per cluster, expandable to 40 nodes virtually, enabling high availability and workload consolidation.
- High Performance Throughput – Delivers up to 7 GB/s read and 1.9 GB/s write per node, with aggregate bandwidth of 238 GB/s per cluster.
- Massive Namespace Scalability – Supports up to 4,160 billion files per namespace and 20,000 shares/10,000 exports, addressing large-scale unstructured data needs.
- Multiprotocol Access – Native support for SMB, NFS, FTP, iSCSI, and S3 tiering to object storage for hybrid cloud integration.

To learn more about VSP One File, reach out to your Hitachi account team or visit <https://www.hitachivantara.com/>.

VSP One Object

Hitachi Virtual Storage Platform One Object is a next-generation, object storage platform designed for large-scale, AI-ready data lakehouse environments. It provides native Amazon S3 compatibility with S3 Table and Apache Iceberg support, enabling in-place analytics and ACID SQL queries directly on object data without requiring transformation or migration. Built on a microservices-based architecture, it supports both hyperconverged and disaggregated deployments, with independent scaling of compute and storage to optimize performance and cost.

The platform delivers high throughput and low latency using NVMe flash, HDD, and QLC SSD media options, making it suitable for AI/ML pipelines, big data analytics, and high-ingest workloads. Enterprise-grade data protection features include erasure coding, immutable object locking, legal hold, asynchronous replication, and automated PII discovery, ensuring compliance and cyber resilience. By unifying unstructured data management with intelligent services – such as event-based lifecycle policies, metadata enrichment, and integrated observability – VSP One Object solves the challenges of explosive unstructured data growth, inefficient siloed storage, and compliance risk, while delivering sustainable efficiency with a 50–70% reduction in rack space and up to 50% lower power consumption compared to competing platforms.



VSP One Object Capabilities:

- Native Amazon S3 Compatibility – Full S3 API support with S3 Tables and Apache Iceberg integration, enabling ACID SQL queries and in-place analytics directly on object data.
- Flexible Media Options – Supports HDD, QLC SSD, and NVMe flash for tiered performance and cost optimization.
- High Performance – Designed for high-throughput ingest and low-latency access, supporting AI/ML training, inference, big data analytics, and real-time pipelines.
- Massive Object Scale – Stores up to 1.25 billion objects per node, scaling to exabyte-class capacity.

For more information about VSP One Object, reach out to your Hitachi account team or visit <https://www.hitachivantara.com/>.

Solution components

This section provides a detailed breakdown of the hardware, software, and networking components validated in the Hitachi Data Lakehouse reference architecture. This section details the essential components that work together to deliver simplified governance for modern data workloads. By understanding each element's role, organizations can confidently deploy a robust platform tailored to their business needs.

Hardware components

The following table lists the hardware components tested in this reference architecture.

- Compute

Vendor	Hardware	Detail Description	Version	Quantity
Hitachi Vantara	Hitachi Advanced HA815 G3 Server	OpenShift control plane and worker nodes CPU: 2 × AMD EPYC 9355 32-Core Processor 512 GB with 16 × 32 GB DDR5-4800 480 GB NVMe SSD Boot Drive 2 × 7.68 TB SSD INT E810 100 GbE 2p QSFP28 PCIe Adapter Dual port BCM 57414 10/25 GbE SFP28 OCP3 Adapter	SPV: 7.42 OS: Red Hat Enterprise Linux 9.4	5

- Storage

Vendor	Hardware	Detail Description	Version	Quantity
Hitachi Vantara	VSP One Block	Block Storage for RDBMS workloads	A3-04-01-40/00 HA-DKC-04-01XR	1
Hitachi Vantara	VSP One File	File storage for semi-structured data and Zetaris metadata, data and user files.	15.5.8424.05	1
Hitachi Vantara	VSP One Object	Object storage for unstructured data and structured HADB parquet data	V3.1.0	1

- Networking

- Storage network:
 - FC SAN switch connections for high-throughput data access between B26 storage and compute nodes
- Data network:
 - High Speed 10/25/100 GbE Ethernet switches for high-throughput data access
- Management network:
 - 1/10GbE switches for OOB and in-band management
- Connectivity options:
 - NFS for file storage
 - S3 protocol for object storage

- Switch hardware

Vendor	Hardware	Detail Description	Version	Quantity
Cisco	Nexus 93180YC-FX/EX (data)	48 × 10/25GbE, 6 × 40/100GbE	NXOS 9.3(13), 10.3(7), & 10.4(3)	2
Cisco	Nexus 92348 (mgmt)	48 × 1GbE, 4 × 1/10/25 GbE, 2 × 40/100 GbE	NXOS 9.3.5	1
Brocade	G720	48 × 16/32 Gbps FC	9.1.1c	2

Software components

The following table lists the key software components that were tested in this reference architecture.

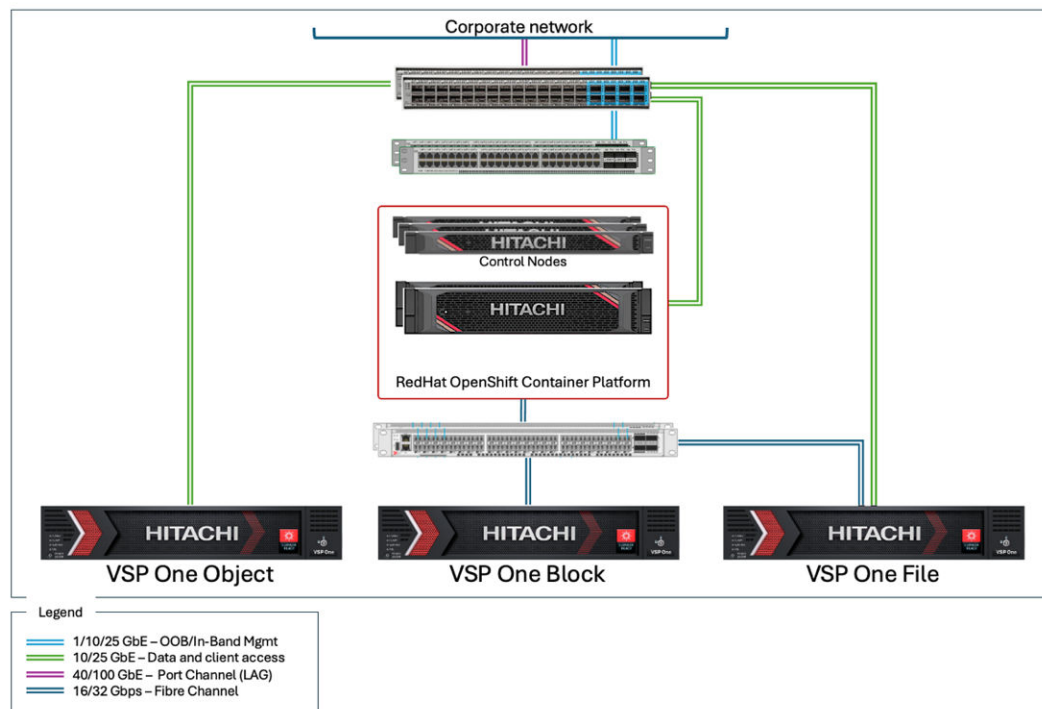
Software	Version	Function
Red Hat Enterprise Linux	9.4	Base OS for OpenShift
OpenShift (OCP)	4.17.27	Container orchestration platform
Zetaris Enterprise	2.4.1	Federated query engine and data virtualization
PostgreSQL	15	Enterprise Data OLAP database engine: Application data source for federated queries
Hitachi Advanced Database (HADB)	6.0	High-speed columnar database for archive data
HSPC CSI Driver	3.16.0	Block storage integration
NAS CSI Driver	4.11.0	File storage integration

This technology stack ensures seamless ingestion, processing, and querying of data across heterogeneous sources, enabling real-time insights and simplified governance for enterprise workloads.

Solution design

This section provides a description of an example Hitachi Data Lakehouse solution architecture, including the physical and logical design, and network topology. Drawing on best practices validated in other Hitachi reference architectures, the design is presented in a modular and scalable manner to accommodate a range of deployment sizes and performance requirements.

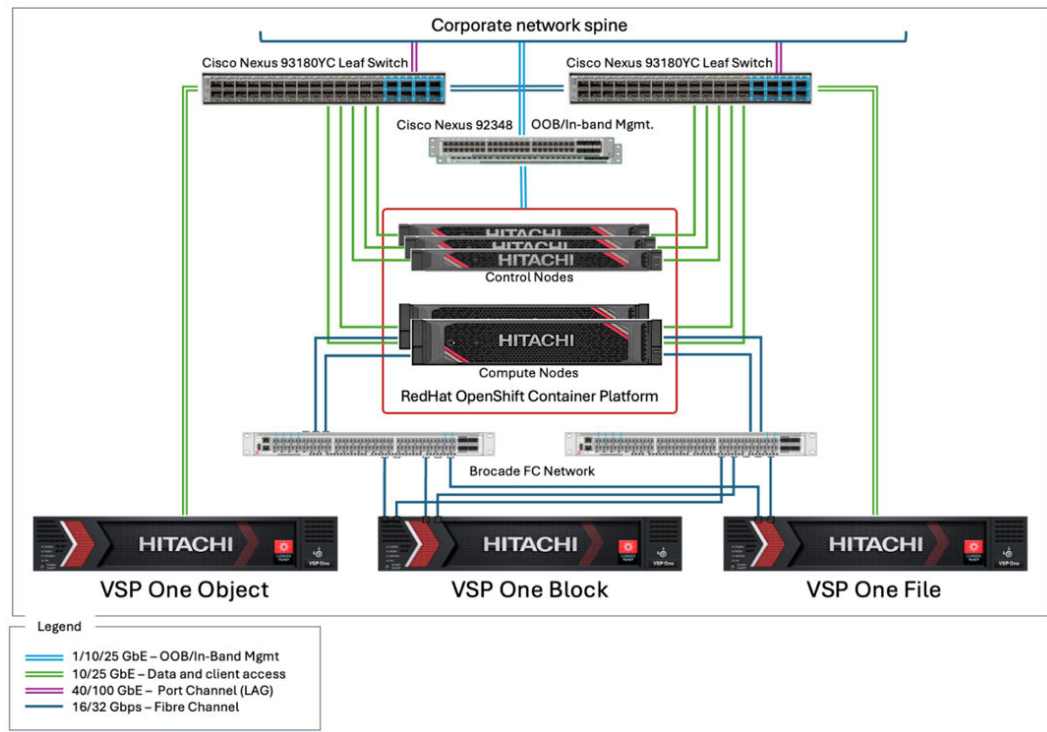
The following are the core building blocks of this architecture:



- **Container Ecosystem:** Dedicated control nodes and worker nodes running Red Hat OpenShift provide the flexibility and scalability required for today's demanding applications. This specific solution design used the Hitachi Advanced Server HA815 G3, however any Hitachi Advanced Server can be used, assuming it supports the Red Hat OpenShift Container Platform version used.
- **Block Storage:** High-speed, resilient storage (VSP One Block) for operational databases and real-time metadata.
- **Object Storage:** Durable, cost-effective storage (VSP One Object) for compliance, and unstructured data, including Parquet datasets for in-place analytics.
- **File Storage:** Scalable file services (VSP One File) for collaborative analytics, shared notebooks, logs, and team data.
- **High-Performance Networking:** Non-blocking, low-oversubscription network architecture (Cisco Nexus and Brocade switches) ensures fast, reliable data access across all components.
- **World-Class Software Stack:** Integrated with Zetaris Enterprise for federated queries and a unified semantic layer, and Hitachi Advanced Database (HADB) for high-performance, in-place SQL analytics on object storage—delivering unified analytics, robust governance, and seamless access to both current and historical data across all storage tiers.

Cisco Nexus switches provide high-speed connectivity for both data access and out-of-band/in-band management, while Hitachi Advanced Server nodes serve as controller and worker nodes for the Red Hat OpenShift Container Platform cluster, delivering scalable and resilient compute resources. These nodes are interconnected via Ethernet and Fiber Channel networks, with Brocade FC switches enabling high-performance storage connectivity. Red Hat OpenShift Container Platform orchestrates, manages, and scales containerized applications such as Zetaris, HADB, and Postgres across the cluster.

At the storage layer, VSP One Object, VSP One Block, and VSP One File systems deliver unified, enterprise-grade storage services for block, file, and object workloads. This modular and layered design supports robust, scalable, and secure operations, and serves as a blueprint for implementing the solution in real-world environments, ensuring that organizations can achieve optimal scalability, availability, and operational efficiency while meeting their business and technical objectives.



Solution Architecture

The physical architecture of the Hitachi Data Lakehouse Solution is designed for modularity, scalability, and high performance, supporting a range of enterprise data workloads. At the compute layer, Red Hat OpenShift Container Platform (RHOCP) is deployed on Hitachi Advanced Server HA815 G3 nodes, forming a cluster with three control nodes and two worker nodes. This cluster orchestrates and manages containerized applications such as Zetaris, HADB, and Postgres, providing resilient and scalable compute resources.

Red Hat OpenShift Container Platform (OCP) Cluster Configuration:

- 3 × Controller nodes (Hitachi Advanced Server HA815 G3)
- 2 × Worker nodes (Hitachi Advanced Server HA815 G3)

Networking is managed by two Cisco Nexus switches:

- Cisco Nexus 93180YC delivers front-end high-speed data access for application and client connectivity and provides high-speed connectivity to VSP One File and VSP One Object storage systems, ensuring low-latency access across the environment.
- Cisco Nexus 92348 supports out-of-band (OOB) and in-band management, ensuring secure and reliable administrative operations.

Storage layer

The storage layer consists of three VSP One platforms, each optimized for specific data availability types. High-performance block storage connectivity is enabled by Brocade G720 Fibre Channel switches, that link the compute nodes to the block storage systems with low latency and high throughput. VSP One File consumes block storage from the Fibre Channel switches, while serving file-based data to lakehouse applications via the front-end Cisco network. VSP One Object connects via the Cisco front-end network to serve object data for active analytics and data preparation as well as serving data for compliance and archival data use cases.

- VSP One Block for RDBMS and OLAP workloads needing more performance and throughput. Robust block storage for structured data.
- VSP One File for semi-structured data such as collaborative workloads, shared notebooks. Scalable file-based storage, allowing for simultaneous data access.
- VSP One Object offering flexible and scalable object storage for structured and unstructured data for active workloads as well as historical, compliance, and log data.
- Brocade G720 Fibre Channel switches provide high-performance storage connectivity between the VSP One Block and the Red Hat OpenShift worker nodes. Alternatively, traditional ethernet switches can be used for block storage connectivity supporting iSCSI or NVMe-over-TCP, offering additional flexibility for environments where Fibre Channel is not preferred or available.
- Depending on the block storage configuration and the number of worker nodes, worker nodes can be direct attached to the block storage. Note that block direct attach will limit the scalability of the solution.

Storage Type	CSI Driver	Access Mode	Purpose
Enterprise Block	HSPC Block CSI Driver	RWO	Operational databases, real-time metadata
Shared NFS	K8S Native NFS CSI Driver	RWX	Collaborative workloads, shared notebooks and data
S3 Object Storage	None: S3 protocol	RWO	Object analytics/Iceberg, active analytics, historical data, compliance archives, logs and caching, archiving

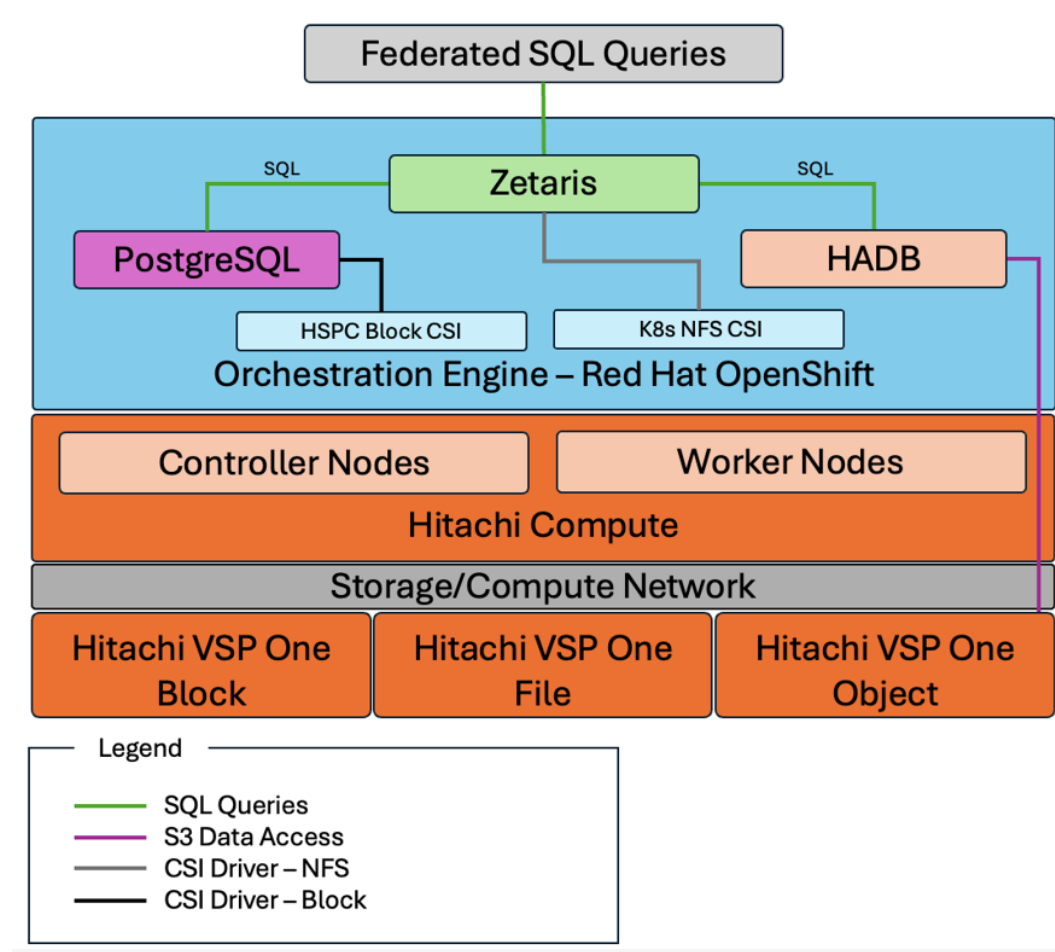
Software layer

Zetaris Enterprise is deployed in its own namespace and runs as a set of pods providing federated query services, a semantic layer, and user-facing interfaces. Zetaris leverages multiple CSI-backed volumes: block storage for metadata workloads, NFS for shared content and general-purpose distributed storage for logs and AI/ML workloads.

- Zetaris metadata is stored on VSP One Block Storage via the Hitachi Block CSI driver.
- Zetaris uses VSP One File storage for shared services such as notebooks, logs, and content storage via the native K8S NFS CSI driver.

PostgreSQL serves as an operational data source that consumes block storage via persistent storage volumes on the VSP One Block to store system catalogs, user data, and operational metadata. Datasets from PostgreSQL are joined with other data from HADB and VSP One File and served through Zetaris. It also uses persistent volumes on VSP One Block sized and tuned for IOPS/latency.

Hitachi Advanced Database (HADB) is configured for multi-modal storage access with direct S3 integration to VSP One Object. It uses ephemeral local storage for spool operations, while all durable data is stored externally in an S3 bucket. HADB serves as a data source for historical and archive data for compliance and archiving use cases.



Solution validation

To validate the Hitachi Data Lakehouse Solution, we used the industry-standard TPC-H dataset to test data access and query federation across multiple storage types and engines.

Dataset Placement:

- The TPC-H dataset was loaded into PostgreSQL (representing active relational data on block storage).
- The same dataset was exported as Parquet files and placed on VSP One File for file-based access.
- Parquet files were also stored on VSP One Object (object storage), where they could be queried via Hitachi Advanced Database (HADB).

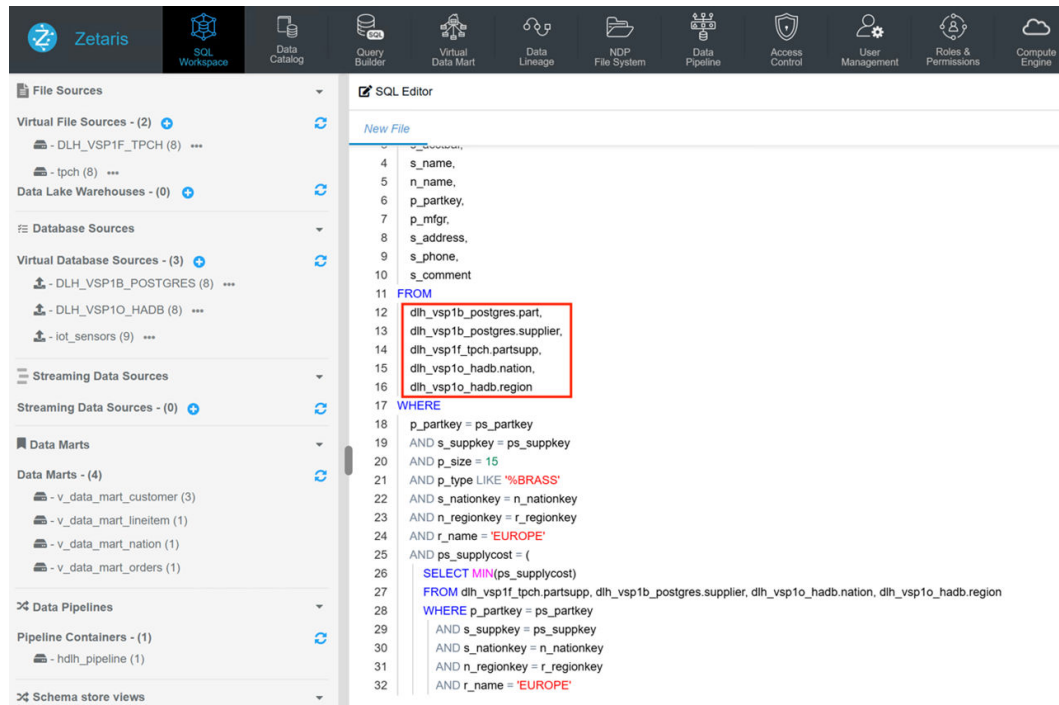
Validation Approach:

- All 22 TPC-H benchmark queries were executed using Zetaris against each individual data source: PostgreSQL, Parquet on VSP1F, and Parquet on Object via HADB. This confirmed that Zetaris could connect to and query each source independently.
- The queries were then modified to create federated queries that joined tables across different data sources (for example, combining data from PostgreSQL and Parquet files). Queries that only referenced a single table were excluded from the federated tests.
- Both individual and federated queries were run using the Zetaris User Interface, as well as through the Zetaris REST API and JDBC connections, to verify consistent behavior across access methods.
- To simulate archival use cases, data was migrated from PostgreSQL into parquet files within VSP One Object buckets. Using HADB, foreign server and foreign tables were created then queried via Zetaris. Federated queries were executed to simulate compliance use cases where archival data might be important for financial auditing purposes.

Test cases

Area	Priority	Description
Functional Testing	P1	Validate connectivity, drivers, and component health
Unified Info Access	P1	Ingest and query data from block, file, and object stores via Zetaris
Data Archival/Retrieval	P2	Move OLAP data to Parquet on object store and query via HADB + Zetaris

The following image demonstrates how data professionals can utilize the Zetaris SQL Workspace to craft federated queries across multiple data sources. In this example, the query federates tables across PostgreSQL on Hitachi block, parquet data in Hitachi file and Hitachi Advanced Database consuming archive data from an S3 bucket on Hitachi object storage.



This validation demonstrates that the Hitachi Data Lakehouse Solution can provide unified, federated query access to data stored across diverse sources, using Zetaris as the semantic and query engine. The solution supports both interactive and programmatic query workflows, and highlights the importance of ongoing interoperability improvements for complex query patterns.

Use case discussion

The Hitachi Data Lakehouse with Zetaris Enterprise is designed to address a set of recurring data challenges: unifying access to heterogeneous stores without heavy ETL, balancing “hot” performance with “cold” economics, enabling collaborative analytics, and maintaining strong governance at scale. The following discussion examines how these needs are met on this reference architecture and highlights operational considerations that matter when moving from proof of concept to production.

Unified analytics without data movement

Enterprises often keep their main business data—like orders or customer records—in databases, while related information such as contracts, logs, or images is stored separately in file or object storage. Traditionally, bringing all this data together for analysis meant copying it into a single warehouse, which is slow, expensive, and creates lots of extra copies to manage.

With this architecture, Zetaris Enterprise acts as a smart “traffic controller” that lets you run a single query across all your data—whether it’s in PostgreSQL (for active records), VSP One File (for shared files), or HADB paired with VSP One Object (for long-term or unstructured data)—without having to move it first. Zetaris breaks each query into parts, sends them to the right storage system, and then combines the results for your analytics tools or dashboards.

Here's how the storage works in practice:

- Databases use fast block storage (VSP One Block) for reliable, high-speed access.
- Shared files and collaborative work use file storage (VSP One File) that allows multiple users or applications to access the same data at once.
- Long-term or unstructured data (like logs, images, or archives) is kept in object storage (VSP One Object). Hitachi Advanced Database (HADB) manages, and queries columnar data stored in parquet files stored on S3 Object Storage, making it easy to analyze large historical datasets.

This setup means you always get the freshest data straight from the source, with fewer data copies to secure and far less complexity in your data pipelines. In short: each type of data stays where it works best, and Hitachi Data Lakehouse with Zetaris lets you analyze everything together—simply and securely.

Lifecycle tiering with in place historical analytics

When you need fast access to current data—like recent transactions or active records—it makes sense to keep that data on high-speed block storage (using SSDs), which is what PostgreSQL on VSP One Block provides. As data gets older and is used less often, it's more efficient to move it to a lower-cost storage tier.

In this solution, older data is exported from PostgreSQL into Parquet files and stored on VSP One Object (object storage). Instead of having to move this data back into a database to analyze it, we use Hitachi Advanced Database (HADB), which can read and query those Parquet files directly from object storage.

Hitachi Data Lakehouse brings it all together by letting users run a single query that combines both the operational and real-time data in PostgreSQL and archival data managed by HADB on object storage. This means users see one unified dataset, no matter where the data lives, and the system automatically uses the most cost-effective storage for each type of data.

If your organization uses tools to automatically move data between storage tiers (like policy-driven data lifecycle management), this approach still works—the queries will seamlessly span both active and archived data. The result: you save on expensive storage for older data but can still analyze everything quickly when you need it.

Collaborative analytics and data science on shared services

Data science teams often work together using shared notebooks and files, and they need easy, reliable access to the data they trust. In this solution, OpenShift makes it possible for multiple users and applications to read and write to the same files at the same time—so everyone can collaborate without running into access issues.

To support this, shared resources such as notebooks, search indexes, and other files are stored on VSP One File using a setup that allows many users or applications to access them at once. Zetaris provides a secure, unified view of all the data—whether it's in PostgreSQL, in files, or stored as objects—so teams can analyze everything in one place.

This setup means teams can experiment and share results quickly, without having to make extra copies of big datasets. It also makes it easier for administrators to manage and secure the data, since each type of data is stored in the place that fits it best. The result: faster collaboration, fewer unnecessary data copies, and a simpler, more reliable way to reproduce results.

Governance, retention, and audit ready analytics

Regulatory programs require durable retention, immutability, and the ability to prove who accessed what, when. VSP One Object contributes object lock and legal hold for non-rewriteable retention, as well as features that support auditability, while still exposing the same data to analytics engines. HADB can read retained Parquet sets via S3 external tables, and Zetaris can restrict exposure through role based semantics so that auditors and compliance users see only the views they are entitled to. Because data remains in place, organizations can reduce the number of derivative datasets and associated risk. Practically, operators should align bucket level retention policies with data classification, document lineage in the Zetaris catalog, and use immutable snapshots/tiering to defend against tampering—all without removing data from its authoritative store.

Operational observability with business context

Troubleshooting and planning for future growth are much easier when you can connect system logs and performance data to real business activities. In this solution, logs and telemetry from applications and the platform can be stored on VSP One File for easy sharing and analysis, or on VSP One Object for long-term, reliable storage. Meanwhile, business transactions and records are kept in your main database system.

Zetaris makes it possible to view and analyze all these different types of data together. For example, if there's a spike in errors, operations teams can quickly see which customers or orders were affected—without having to move data into a separate log management system.

Each type of data is stored in the place that fits it best:

- Shared logs go on file storage for easy access and collaboration.
- Business records stay in high-performance databases.
- Long-term logs and archives are kept in object storage for durability and cost savings.

This approach keeps analytics consistent and efficient, even as data grows and ages over time.

How the patterns generalize on this reference design

Throughout this architecture, a few important principles come up repeatedly. Different types of data are managed separately for good reason. Zetaris keeps track of its own system information and operations in one area, while active business data—such as transactions or customer records—can be stored in whichever database best fits your organization's needs. Each type of data should be managed and backed up on its own to ensure high performance and reliable protection.

Services that need to be shared by multiple users or applications—like search indexes, collaborative notebooks, or shared files—are stored on VSP One File for easy sharing and teamwork. This makes it simple to scale up and support more users without running into bottlenecks. Meanwhile, older data or information that doesn't change much, such as historical records, logs, or large collections of documents, is kept in VSP One Object. This storage is ideal for long-term retention and makes it easy to analyze data using tools like HADB, without needing to move it elsewhere first.

Hitachi Data Lakehouse brings everything together by acting as the control center for your data. It allows users to access and analyze all their information in one place, with the right security and governance in place. This ensures that people can trust the data they're working with and always get up-to-date results. By following these principles, the architecture keeps each type of data in the storage that suits it best, while providing users with a single, secure, and easy way to access everything they need.

Measuring success

Organizations usually measure the success of their data platform in three main ways. First, they look at how quickly they can get answers from their data—since queries run directly on the latest information, there's no need to wait for overnight data processing. Second, they value simplicity: with fewer data pipelines and copies to manage, and clearer backup routines for important systems, day-to-day operations become much easier. Third, they consider cost efficiency: fast, expensive storage is used only for active data, while older information is stored more affordably while remaining easy to analyze.

In everyday use, teams also notice that their key performance indicators (KPIs) become more consistent once everyone uses the same unified data layer for business intelligence and data science. These benefits are possible because the platform is designed to connect and analyze data where it already lives, rather than forcing everything into one big warehouse. At the same time, it meets enterprise standards for security, reliability, and support—whether running on OpenShift or VSP One storage.

Conclusion

The Hitachi Data Lakehouse with Zetaris Enterprise delivers a modern, unified platform for managing and analyzing data across the entire organization. By combining the strengths of Hitachi's VSP One storage portfolio with Zetaris' federated query engine and semantic layer, this solution makes it possible to access, analyze, and govern data wherever it lives—whether in databases, files, or object storage—without the need for complex data movement or duplication.

This architecture is designed for real-world flexibility. It supports fast, reliable access to active business data, easy collaboration on shared files, and cost-effective long-term storage for historical or unstructured information. With tools like Hitachi Advanced Database (HADB), organizations can run powerful analytics directly on large datasets stored in object storage, making it easier to unlock insights from both current and archived data.

Operationally, the platform simplifies day-to-day management by reducing the number of data pipelines and copies, clarifying backup routines, and keeping each type of data in the storage tier that fits it best. Teams benefit from faster access to fresh data, more consistent reporting, and a single, secure way to work with information across the enterprise.

Most importantly, this solution is built to grow and adapt. It honors enterprise standards for security, reliability, and support, and is ready for future expansion as data volumes and business needs evolve. By favoring in-place analytics and federation over consolidation, organizations can achieve better performance, lower costs, and greater agility—while maintaining full control and governance over their data.

In summary, the Hitachi Data Lakehouse with Zetaris Enterprise reference architecture provides a solid foundation for organizations looking to modernize their data platforms, accelerate analytics, and drive smarter business decisions—today and into the future.

Hitachi Vantara



Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA

HitachiVantara.com/contact