VSP One Object with On-Premise Native S3 Tables for Data Lakehouse, AI, and Analytic Workloads

MK-SL-386-00
November 2025

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.

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.	November 2025

Reference Architecture and Feature Description

This paper describes the following:

- VSP One Object on-premises solution for data lakehouse, AI and Analytic workloads
- VSP One Object with Apache Iceberg catalog integration and on-premises native Amazon S3 tables support

Unlocking timely insights with VSP One Object

At its core, data analytics is about delivering the right information to the right people—or increasingly, to AI agents—at precisely the right moment. While the goal is simple to articulate, achieving it has long been a complex challenge.

With the latest release of Virtual Storage Platform One (VSP One) Object, that challenge is being redefined. This breakthrough platform introduces a powerful foundation for modern data lakehouse, analytics and AI-driven workloads. By combining scalable object storage with support for open table formats and enriched metadata capabilities, VSP One Object empowers organizations to harness the full potential of their data. It also lays the groundwork for seamless integration with generative AI, enabling smarter, faster, and more context-aware decision-making.

The evolution of data infrastructure: From warehouses to object stores

Not long ago, building a data warehouse was a slow, IT-driven process. Business users defined requirements, but IT teams managed everything—from data transformation to dashboard design. Even with the rise of self-service BI tools, users still depended on curated datasets, which took time to prepare and delayed access to insights.

To speed things up, organizations began storing raw data in large-scale data lakes, deferring transformation and cleansing. Technologies like Hadoop and MapReduce enabled this shift, allowing companies to collect massive volumes of data under the belief that more data meant more value. Apache Spark became the dominant solution for data processing at scale. But raw data, like crude oil, must be refined to be useful. Data lakes lacked governance and structure, making them less effective for business users.

For data scientists, however, data lakes offered flexibility for building machine learning models. This divergence in needs—raw data for AI, curated data for business—highlighted the limitations of Hadoop-based architectures, which tightly coupled compute and storage, making them hard to scale.

The complexity of managing Hadoop clusters was eventually eliminated with the advent of object stores. It turned out that many workloads, such as Apache Spark jobs, would work equally well using AWS S3-compatible object stores. S3-compatible object stores, decoupled compute from storage, simplifying scalability and operations.

Metadata: The key to unlocking value in data lakes

While data lakes built on object stores are ideal for storing both structured and unstructured data, they have not been without their challenges. Without proper oversight, these environments can quickly become “data swamps”—filled with redundant, obsolete, or trivial (ROT) data that clutters storage and obscures valuable insights.

Another issue is dark data—information that has been collected, processed, and stored but never used. It consumes resources without delivering business value. Often, potentially useful data remains hidden simply because we do not know what is in it or where to find it. This creates risks, especially when sensitive or personally identifiable information (PII) is buried in unclassified documents.

The solution lies in metadata. By identifying what objects exist and associating rich metadata with them, organizations can classify, protect, and manage data more effectively. Metadata enables searchability, governance, and access control—turning a chaotic data lake into a well-organized, insight-ready resource.

Open table Formats: Bringing structure and consistency to object stores

After object storage, the next key ingredient for delivering timely insights is the open table format. While organizations have long stored structured data in formats like CSV, JSON, ORC, Avro, and especially Parquet, managing these files at scale presents challenges. Parquet, now the standard for storing structured data in object stores, is immutable making it ideal for governance but difficult or near impossible to update. As a result, new Parquet files are continually added, forming logical datasets that are hard to query consistently.

Without a unified view, querying many small files becomes inefficient. To address this, files are regularly compacted into larger ones, improving performance and reducing storage costs. Historically, this process required manual effort.

What if we could treat a collection of Parquet files as a logical table, like a database? Better yet, what if we could apply ACID transactional guarantees—ensuring consistency even when multiple processes read and write simultaneously?

This is exactly what open table formats like Apache Iceberg enable. Iceberg organizes Parquet files into logical tables and maintains metadata about schema, snapshots, and file locations. Clients like Apache Spark interact with an Iceberg catalog to read and write data reliably, ensuring consistency and enabling rollback to previous snapshots if needed.

These capabilities bring database-like functionality to object stores, forming the foundation of the data lakehouse—a hybrid architecture that combines the scalability of data lakes with the structure and reliability of data warehouses.

In December 2024, AWS introduced S3 Tables, a native implementation of Iceberg tables within S3. Stored in dedicated “table buckets,” S3 Tables offer optimizations and optional automation for tasks like snapshot expiration and file compaction. While they require AWS Glue as the catalog and use flat namespaces, they remain fully compatible with Iceberg clients, preserving openness and avoiding vendor lock-in.

Native support of S3 Tables in VSP One Object

On August 12, 2025, Hitachi Vantara introduced native support for S3 Tables and table buckets in VSP One Object, becoming the first vendor to offer these capabilities for on-premises object storage. VSP One Object includes full S3 Table APIs and an embedded Iceberg REST catalog, tested with Apache Spark and Trino. Compatible with any Iceberg client supporting formats v1 and v2, it also offers optional automated table management. By integrating directly with the catalog, VSP One Object tracks all S3/Iceberg tables—bringing powerful lakehouse capabilities to enterprise environments.

VSP One Object extends its support for S3 Tables by enabling native querying through BI tools. To simplify integration with SQL-based tools, VSP One Object maps table buckets, namespaces, and tables to SQL catalogs, schemas, and tables, respectively. This flat namespace model mirrors AWS's approach and enables seamless compatibility.

For small to medium-sized tables, BI tools offer a convenient way to access data.

Using the Apache Arrow Flight SQL endpoint, business users can connect via standard JDBC drivers and query S3 Tables directly—no additional configuration required. Similarly, Python client libraries (for example, SQL Alchemy) provide access to S3 Tables within Jupyter notebooks for data analysis via Pandas or to train machine learning models.

For larger datasets or complex workloads, we recommend using optimized Iceberg clients—such as Apache Spark for ETL and Trino for low-latency SQL queries.

Engine-agnostic lakehouse architecture with VSP One Object

VSP One Object includes built-in support for S3 Table APIs and an embedded Iceberg REST catalog, making it an ideal foundation for on-premises data lakehouse. A key advantage of the Iceberg format is its engine independence—multiple clients can read and write to the same table concurrently, with ACID guarantees ensuring data consistency through optimistic concurrency control.

This flexibility encourages the use of multiple engines. Apache Spark excels at writing and transforming data in Iceberg tables, especially for batch processing. Trino, on the other hand, offers low-latency SQL query execution, making it ideal for interactive analytics at scale.

To simplify deployment, Hitachi Vantara provides a reference architecture and configuration guidance for integrating Spark and Trino with VSP One Object. This enables organizations to build a robust, scalable lakehouse environment that supports diverse workloads—from ETL pipelines to real-time business intelligence—while maintaining consistency, performance, and openness.

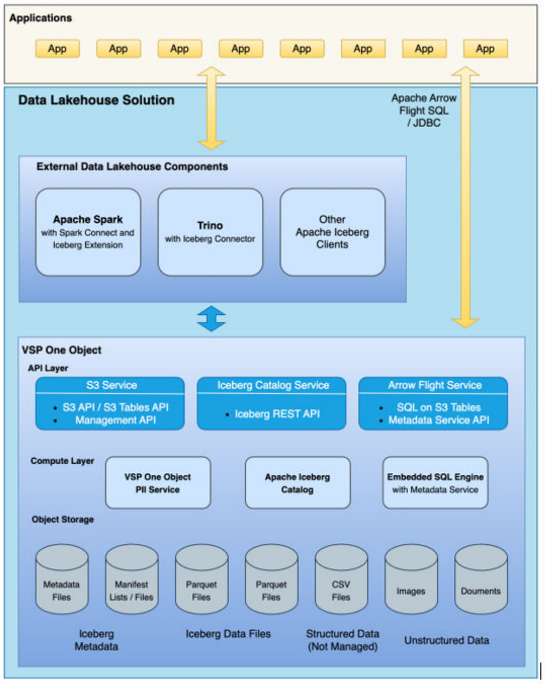
Conceptually, the configuration on an Iceberg client involves the following steps (details can be found in the product documentation).

- Apache Spark
  - Add the following `spark_jars_packages`:
    - `org.apache.iceberg:iceberg-spark-runtime`
    - `org.apache.iceberg:iceberg-aws-bundle`
  - Configure the connection to the Iceberg REST service in VSP One Object.
    - Set the Spark SQL Catalog to `org.apache.iceberg.spark.SparkCatalog`.
    - Set the catalog URI to the VSP One Object Iceberg REST Catalog endpoint.
    - Specify the catalog name.
    - Set the catalog type to `rest`.
  - Add a trust store configuration that is used for configuring VSP One Object self- signed certificates.
- Trino
  - Set the Iceberg catalog type to `rest`.
  - Set `iceberg.rest-catalog.uri` to the VSP One Object Iceberg REST API endpoint for a specific catalog.
  - Set `iceberg.rest-catalog.warehouse` to the S3 Table Bucket in VSP One Object.
  - Configure authentication.
- JBDC Client Example

  We use DBeaver as an example for a BI client that can query the Arrow Flight SQL endpoint in VSP One Object from JDBC.

  - Download the DBeaver community edition.
  - Download the .JAR file of the official Apache Arrow Flight SQL JDBC Driver.
  - In `DBeaver/Database Menu/Driver Manager/Libraries`, click Add File and add the `Arrow Flight SQL JDBC .JAR` file.
  - Specify the driver settings.
  - Create a new database connection and provide credentials for authentication.

We merely provided a conceptual overview of what is required to connect Apache Iceberg clients as well as BI clients to VSP One Object. Details can be found in the documentation which includes a step-by-step guide for setting up Apache Spark and Trino with VSP One Object to deploy a modern on-premises data lakehouse solution in record time!



Conceptual Data Lakehouse Solution

VSP One Object provides the following APIs:

- VSP One Object Management API
- S3 API
- S3 Tables API
- Iceberg REST API
- Apache Arrow Flight SQL

Apache Iceberg clients, such as Apache Spark and Trino communicate with VSP One Object from the S3 and the Iceberg REST APIs.

Analytic applications can access data in VSP One Object via Iceberg clients or directly via Apache Arrow Flight SQL (including JDBC).

VSP One Object Metadata is exposed as fully managed S3 Tables and can therefore also be queried from Iceberg clients and Arrow Flight SQL.

Real-world flexibility with open table formats

After Apache Spark and Trino are deployed with VSP One Object, the benefits of open table formats become clear. Because S3 Tables are decoupled from query engines, multiple users and tools can access the same data simultaneously—each optimized for their specific task.

Imagine a company analyzing public airline on-time performance data to improve travel planning. The data is stored in an S3 Table within VSP One Object. A business user, using a BI tool connected via the Arrow Flight SQL endpoint, queries the table to compare flight delays for connections through Chicago and Denver—quickly identifying weather-related risks during winter travel.

Another employee uses Trino to run complex, multi-stop route analyses, while a data scientist leverages Apache Spark to transform the same dataset into training sets for a machine learning model predicting future delays.

Each user accesses the same S3 Table, using the engine best suited to their task. This flexibility—enabled by open table formats—eliminates the limitations of one-size-fits-all data platforms and empowers smarter, faster decision-making across the organization.

Smarter metadata management with VSP One Object

Beyond scalable storage and S3 Table support, VSP One Object offers powerful metadata capabilities that enhance data classification, governance, and discoverability. Adding or updating metadata often requires rewriting the entire object, which becomes inefficient—especially when multiple services need to annotate the same data.

Consider a scenario where numerous services enrich object metadata: a personally identifiable information (PII) detection service flags sensitive content, a classification service identifies file types, a language service detects languages, and an ETL pipeline tags data quality levels (e.g., raw, curated, published). Additional services might classify documents by domain (HR, legal, healthcare) or even by medical record type. Storing all these properties together using standard S3 metadata quickly becomes messy and hard to manage.

To solve this, VSP One Object introduces a flexible metadata service (MDS) that allows users to define named groups of properties—each representing a logical set of metadata. These groups can be atomically and efficiently attached to objects, buckets, or S3 Tables without interfering with other metadata. For example, a “PII Metadata” group can be added by a PII service, while other groups remain untouched.

All metadata is stored immutably in S3 Tables, ensuring tamper-proof governance and auditability. These metadata tables are queryable via Iceberg clients or BI tools, enabling powerful cross-group queries. For instance, users can identify all PDF documents containing sensitive data created in the last 24 hours by querying across PII, file type, and standard metadata groups.

VSP One Object also supports user-defined entity types—such as LLMs, embeddings, and their relationships—giving organizations full control over how metadata is structured and applied. Out-of-the-box, it provides historical tracking of object attributes and built-in PII detection, with APIs and an SDP for custom metadata integration.

This approach transforms metadata from a static annotation into a dynamic, queryable asset—enabling smarter data governance, improved searchability, and seamless integration across analytics and AI workflows.

VSP One Object: A unified platform for analytics, AI, and all your data lakehouse requirements

Delivering timely, relevant insights, whether to users or AI agents—requires a platform that can manage structured and unstructured data with scale, security, and flexibility. VSP One Object is purpose-built for this, offering a comprehensive foundation for modern data lakehouse, analytics, and AI workloads.

For structured data, VSP One Object supports Apache Iceberg and on-prem native S3 Tables, enabling Parquet files to be treated like database tables with ACID guarantees, schema evolution, and time travel. It allows BI tools to query small to medium datasets directly without any additional configuration. Integration with Apache Spark and Trino supports everything from batch ETL to interactive analytics, with reference architectures simplifying deployment.

For unstructured data, VSP One Object maintains an immutable log of all objects in S3 Tables, compatible with AWS S3 metadata. Its advanced Metadata Service allows users to define custom schemas and named property groups, enabling precise classification, governance, and search. Metadata is queryable via Iceberg clients or BI tools and integrates easily with external catalogs.

This metadata foundation also powers AI workloads. AI agents can interact with VSP One Object using APIs and SQL-based protocols to execute tasks like identifying sensitive data, reorganizing files, or locating specific documents. Structured metadata makes these tasks accurate and efficient. Techniques like GraphRAG allow LLMs to extract entities and relationships from documents and store them as metadata—enhancing discoverability and trust.

From dashboards to deep learning, VSP One Object delivers a unified, scalable, and intelligent platform for all your analytic and AI needs.

Summary

VSP One Object is the first on-premises object store to support S3 Table APIs, and it includes powerful built-in features like an embedded Apache Iceberg catalog, a zero-config SQL queries, and an advanced metadata service. Whether you are building a Spark- or Trino-based lakehouse, integrating AI workflows, or enabling BI access, VSP One Object delivers scalability, simplicity, and intelligence. It not only streamlines structured data management but also enhances AI applications by making metadata easily accessible—ensuring users and AI agents get the right information, at the right time, with full governance and flexibility.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA

HitachiVantara.com/contact