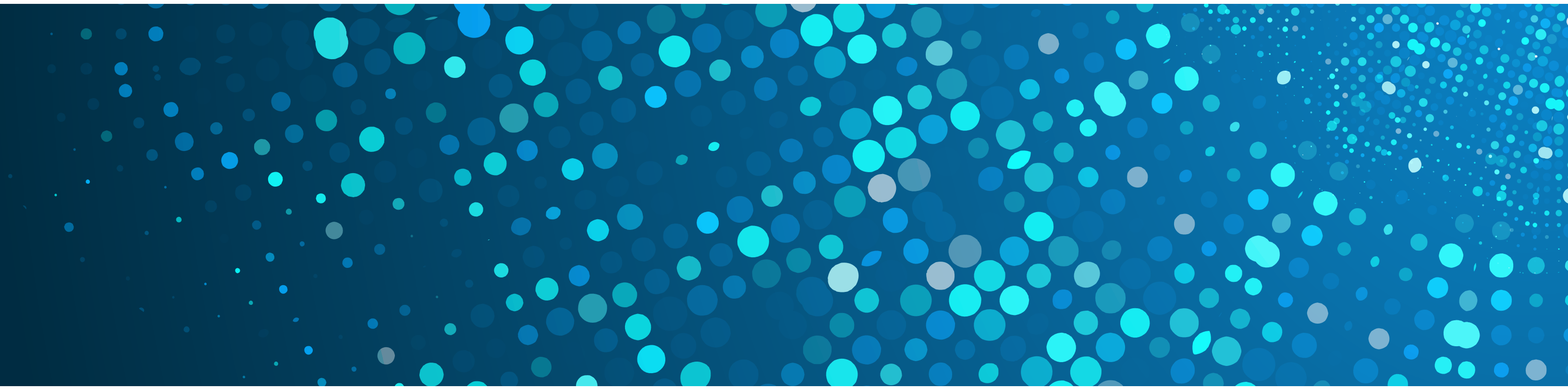
Solution Brief



Cisco UCS Integrated Infrastructure for Big Data and Analytics with Transwarp Data Hub

Bring comprehensive data warehouse capabilities to big data.

Highlights

Proven platform for enterprise data warehouse

- This fifth-generation platform has been deployed across major industries such as agriculture, education, entertainment, finance, healthcare, industrial, insurance, manufacturing, public-sector, service provider, and utilities.
- The solution has been demonstrated through industry-standard benchmarks.

Designed, tested, and validated for faster time to value

- Cisco® Validated Designs facilitate faster, more reliable, and more predictable customer deployments.
- Validated designs provide design, scalability, and performance recommendations.

Built on Cisco UCS foundation

- The Cisco Unified Computing System™ (Cisco UCS®) M5 platform offers complete integration of computing, networking, and storage resources with unified management along with high performance, expandable storage, and scalability for big data systems.
- Cisco UCS is a fabric-centric architecture designed for business acceleration, providing true on-demand infrastructure with a system that grows gracefully and incrementally.

Designed to scale from small to very large as applications demand

- With Cisco Application Centric Infrastructure (Cisco ACI™), you can easily scale a cluster to thousands of nodes.

- Cisco ACI implements an application-aware, policy-based approach that treats the network as a single entity rather than a collection of switches.

Automated deployment and configuration

- Enable one-click provisioning, installation, and configuration of big data infrastructure using Cisco UCS Director Express for Big Data.

Full SQL Standard and Atomicity, Consistency, Isolation, and Durability (ACID) support

- Transwarp Inceptor, a SQL-on-Hadoop engine for batch processing, is the first commercial Hadoop distribution that supports SQL 2003, Oracle PL/SQL, DB2 SQL PL, and ACID and Create, Read, Update, and Delete (CRUD) functions.

  This feature enables third-party tools to conveniently and transparently work with Transwarp Data Hub (TDH) through Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) drivers.

Superior performance with low-latency streaming

- TDH supports complex analytic workloads and fast queries on petabyte-scale data sets.
- TDH also supports event-based streaming with latency as low as 5 milliseconds (ms) and microbatching for large-throughput workloads using Transwarp Slipstream technology.

Cisco and Transwarp provide comprehensive data warehouse capabilities in Hadoop

Big data technology is rapidly transforming modern business. Apache Hadoop is the most mature of the big data technologies available today and has the largest installed base. However, even with the cost benefits of open-source Hadoop, implementing sophisticated scalable data warehouses capable of performing streaming processing at scale remains a challenge. This problem is the result of both a shortage of skills and the inherent complexity of distributed systems.

Transwarp Data Hub (TDH) is specifically designed to address these issues. TDH allows developers to use a single language, SQL, for batch processing, interactive analysis, streaming analytics, and searches. Data in remote heterogeneous infrastructures is available using SQL through Transwarp Inceptor, TDH's analytic engine. Building streaming applications in TDH is straightforward. Using Tranwarp Slipstream technology, developers can use the same SQL language for streaming as they do when accessing a database.

TDH enables significant cost savings without sacrificing performance. Enterprises do not need to spend huge amounts of money to efficiently create a sophisticated, scalable big data system.

Cisco UCS Integrated Infrastructure for Big Data and Analytics

Organizations today must be sure that the underlying physical infrastructure can be deployed, scaled, and managed in a way that is agile enough to adapt as workloads and business requirements change. Cisco UCS® Integrated Infrastructure for Big Data and Analytics has redefined the potential of the data center with its revolutionary approach to managing computing, network, and storage resources to successfully address the business needs of IT innovation and acceleration. This solution provides an end-to-end architecture for processing high volumes of structured and unstructured data for both real-time and archival purposes.

Figure 1 shows Cisco UCS Integrated Infrastructure for Big Data and Analytics with TDH.

Cisco UCS 6300 Series Fabric Interconnects

Cisco UCS 6300 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for servers, with Cisco UCS Manager providing integrated, unified management for all connected devices. The Cisco UCS 6300 Series Fabric Interconnects are a core part of the Cisco Unified Computing System™ (Cisco UCS), providing low-latency, lossless 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and Fibre Channel functions.

Cisco fabric interconnects offer the full active-active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications. Cisco UCS Manager enables rapid and consistent server configuration using service profiles and automates ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS Rack Servers (C240 M5 and C220 M5)

Cisco UCS M5 Rack Servers are dual-socket, 2-Rack-Unit (2RU) servers. They offer industry-leading performance and expandability for a wide range of storage and I/O-intensive infrastructure workloads for big data and analytics.

These servers use the latest Intel® Xeon® Scalable processors, with up to 28 cores per socket. They support up to 24 Double-Data-Rate-4 (DDR4) Dual Inline Memory Modules (DIMMs) for improved performance and lower power consumption.The DIMM slots are 3D XPoint ready, supporting next-generation nonvolatile memory technology.

Depending on the server type, Cisco UCS Rack servers have a range of storage options. The Cisco UCS C240 M5 Rack Server supports up to 24 Small-Form-Factor (SFF) 2.5-inch drives (with support for up to 10 Non-Volatile Memory Express [NVMe] PCI Express [PCIe] Solid-State Disks [SSDs] on the NVMe-optimized chassis version) or 12 Large-Form-Factor (LFF) 3.5-inch drives plus 2 rear hot-swappable SFF drives with a 12-Gbps SAS modular RAID controller. The Cisco UCS C220 M5 Rack Server supports up to 10 SFF 2.5-inch drives (with support for up to 10 NVMe PCIe SSDs on the NVMe-optimized chassis version). Additionally, all servers have two modular M.2 cards that you can use for boot. A modular LAN-On-Motherboard (mLOM) slot supports dual 40 Gigabit Ethernet network connectivity with the Cisco UCS Virtual Interface Card (VIC) 1387.

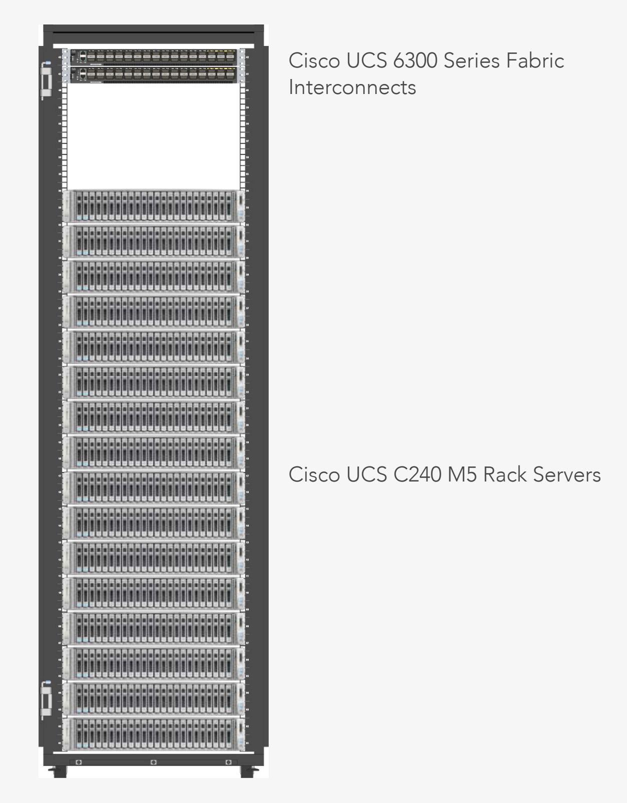

Cisco UCS 6300 Series Fabric Interconnects
Cisco UCS C240 M5 Rack Servers

Figure 1. Cisco UCS Integrated Infrastructure for Big Data and Analytics with Transwarp Data Hub

Transwarp Data Hub

TDH is a big data platform recognized by leading analysts. It provides a number of state-of-the-art technologies that dramatically improve the usability, performance, and stability of the Apache Hadoop platform. It allows enterprises to build core business systems and create new applications in a more efficient, cost-effective manner.

As illustrated in Figure 2, TDH includes six major products for getting value from big data:

- Inceptor: For big data analytics
- Slipstream: For real-time computing
- Discover: For extracting value from data through machine learning
- Hyperbase: For unstructured data processing
- Search: For building enterprise search engines
- Sophon: For deep learning

TDH Release 5.0 adds a big data development toolkit called Transwarp Studio, greatly improving the productivity and efficiency of big data development efforts. Transwarp Studio provides a unified access platform for the following tools:

- Transport: A visualized tool for designing and building Extract, Transform, and Load (ETL) jobs
- Workflow: A service for designing, debugging, controlling, and analyzing workflows in a visual way
- Rubik: A web-based tool for designing Online Analytical Processing (OLAP) “cubes,” which can be materialized in either Hadoop Distributed File System (HDFS) or Transwarp Holodesk
- Governor: A metadata management and data governance tool
- Waterdrop: A SQL Integrated Development Environment (IDE)

In addition, TDH 5.0 implements an innovative architecture that makes each product a service in the cloud using containers and adopts Docker and Kubernetes platforms to help with resource management. These improvements provide many benefits:

- Cluster deployment, installation, and maintenance are significantly simplified
- Multitenant scenario support is improved through flexible and automated resource management through which each tenant can build its own applications as if in an isolated environment

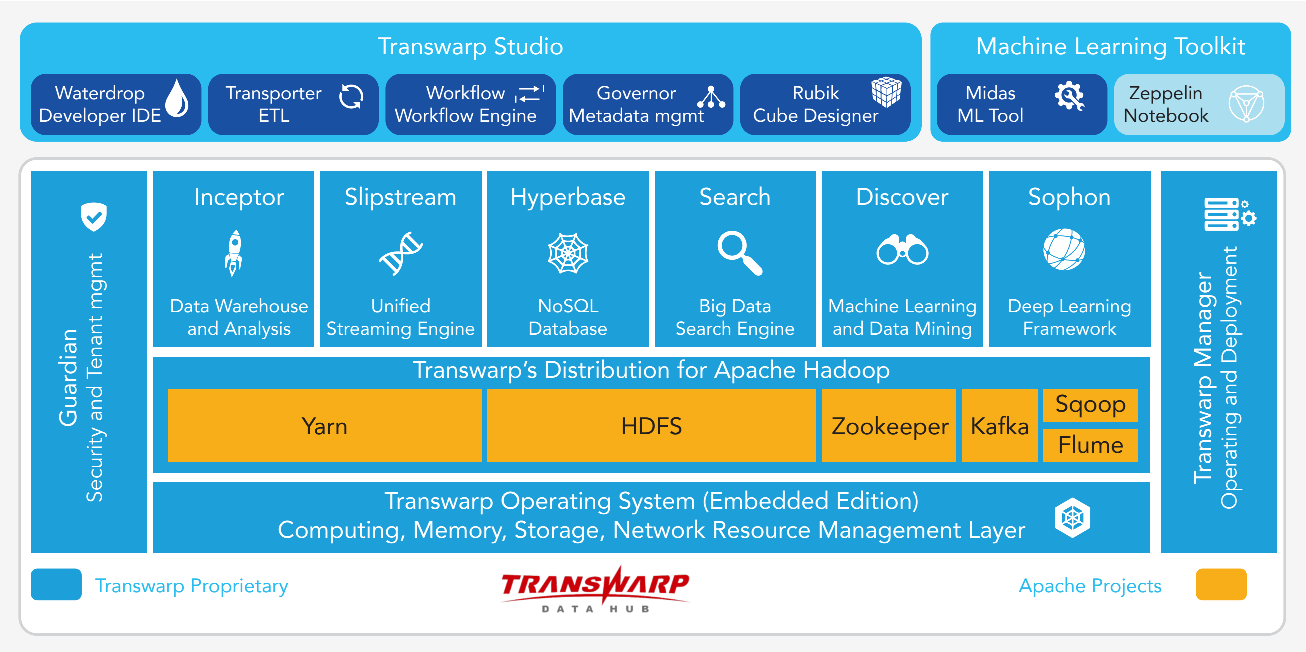


Figure 2. Transwarp Data Hub 5.0

Features of Transwarp Data Hub

TDH offers the following main features:

- **Extreme performance and scalability:** TDH has a highly optimized distributed execution engine, Inceptor, which scales with cluster size and tunes data shuffling and broadcasting logic for better performance. Inceptor employs Holodesk, a columnar storage engine that uses cost-based and rule-based optimization to accelerate reading speed. In addition, by applying the OLAP cube technique, Inceptor generates fast response times for interactive data analysis use cases.
- **Docker and big data platform:** Beginning with Release 5.0, TDH is deployed on the Transwarp Operating System (TOS), a cloud operating system specialized for big data platforms. Based on Docker and Kubernetes, TOS supports one-click installation of TDH. It allows elastic capacity increase and decrease as well as priority-based preemptive-level resource control and fine-grained resource allocation.
- **Full SQL and ACID support:** TDH is the first SQL-on-Hadoop product on the market to provide full support for SQL 2003, Oracle PL/SQL, and DB2 SQL PL at the same time. TDH eases development of streaming applications through support for the StreamSQL standard, including SQL 99 and stream extensions. For transactional applications, Transwarp is the first commercial big data platform to provide ACID support through serializable isolation with Two-Phase Locking (2PL) and Multiversion Concurrency Control (MVCC).
- **Low-latency streaming processing:** Transwarp Stream is the first spark-based event-driven streaming engine. It integrates microbatching and an event-driven engine to support different use scenarios. In microbatching mode, data is processed in microbatches, and a SQL program is run over the batches. This mode enables large-scale throughput, making it suitable for applications such as video detection. In event-driven mode, every event triggers a computing process, allowing the creation of applications that are sensitive to latency, such as fraud-detection applications.
- **Graphical big data development toolkit:** Transwarp Studio is designed to build highly efficient enterprise data warehouses by integrating the access interfaces of Governor, Workflow, Transporter, Rubik, and Waterdrop on a single platform. This approach provides users with a clearer sense of how to choose and operate the tools under any given circumstances, reducing the technical barriers to development of big data systems.
- **Robust machine-learning and deep-learning capabilities:** Transwarp Discover offers machine-learning capabilities for data analysts and scientists through R and Python. Transwarp Sophon provides distributed deep-learning platforms to facilitate the development of Artificial Intelligence (AI) applications.
- **Unified security and multitenant management:** Security control and resource management are centrally maintained by a service called Transwarp Guardian, which provides fine-detailed tenant management for the use of Docker functions.
- **Capability to process a wide variety of data types:** Transwarp Hyperbase can be used to store and compute both structured and unstructured data including logs, JavaScript Object Notation (JSON) and XML, and binary data such as images and videos.
- **Full-text searching on big data:** Transwarp Search helps with the construction of big data search engines within enterprises and provides full-text searches on big data through SQL.
- **Ease of operation and management:** Transwarp Manager is a component for deploying and managing clusters. One-click installation has been supported since 2014. Alerts and health-check features are implemented to reduce the overhead of managing clusters. All components of TDH are optimized for Docker, and the computing engines also use Kubernetes for resource management.

Reference architecture for TDH

The Cisco and Tranwarp reference architectures fir TDH are optimally designed and tested to help ensure a balance between performance and capacity. These configurations can be deployed as-is or used as templates for building custom configurations. The solution can be customized based on workload demands, including expansion to thousands of servers through the use of Cisco Nexus® 9000 Series Switches. With its blazingly fast computing and memory and flexible storage options, this next-generation infrastructure can be used to power extremely fast data access in the large-capacity storage required for modern applications.

Table 1 lists the performance and capacity options for the Cisco UCS Integrated Infrastructure for Big Data and Analytics.

Table 1. Cisco UCS Integrated Infrastructure for Big Data and Analytics Options

Bundle	Performance	Capacity	High Capacity
Server SKU	UCS-SP-C240M5-A2	UCS-SPC240M5L-S1	UCSS-SP-S3260-BV
Servers	16 x Cisco UCS C240 M5 with SFF drives	16 x Cisco UCS C240 M5 with LFF drives	8 x Cisco UCS S3260 Storage Server, each server node with:
CPU	2 Intel Xeon Processor Scalable Family 6132 (2 x 14 cores, 2.6 GHz)	2 Intel Xeon Processor Scalable Family 4110 (2 x 8 cores, 2.1 GHz)	2 Intel Xeon processor E5-2680 v4 CPUs (2 x 14 cores, 2.4 GHz)
Memory	12 x 16 GB 2666 MHz (192 GB)	12 x 16 GB 2666 MHz (192 GB)	8 x 32 GB 2400 MHz (256 GB)
Boot	M.2 with 2 x 480-GB SSD	M.2 with 2 x 480-GB SSD	2 x 480-GB enterprise value boot SSD
Storage	26 x 1.8TB 10K rpm SFF SAS HDDs or 12 x 1.6 TB Enterprise Value SATA SSDs.	12 x 8TB 7.2K rpm LFF SAS HDDs + 2 SFF rear hot-swappable 1.6TB Enterprise Value SATA SSDs	24 x 6 TB 7.2K rpm LFF SAS HDDs
VIC	40 Gigabit Ethernet (VIC 1387)	40 Gigabit Ethernet (VIC 1387)	40 Gigabit Ethernet (VIC 1387)
Storage Controller	Cisco 12-Gbps SAS modular RAID controller with 4-GB Flash-Based Write Cache (FBWC) or Cisco 12-Gbps modular SAS Host Bus Adapter (HBA)	Cisco 12-Gbps SAS modular RAID controller with 2-GB Flash-Based Write Cache (FBWC) or Cisco 12-Gbps modular SAS HBA	Cisco 12-Gbps SAS modular RAID controller with 4-GB FBWC
Network Connectivity	Cisco UCS 6332 Fabric Interconnect	Cisco UCS 6332 Fabric Interconnect	Cisco UCS 6332 Fabric Interconnect

Note: For the management nodes, use three Cisco UCS C240 M5 Rack Servers, each with two Intel Xeon Scalable processors 384 GB of memory, a 12-Gbps SAS RAID controller with a 4-GB cache, 10 x 1.8-TB 10,000-rpm SFF SAS drives, and Cisco UCS VIC 1387 (two 40 Gigabit Ethernet Quad Small Form-Factor Pluggable [QSFP] interfaces).

Conclusion

The fifth generation of Cisco UCS Integrated Infrastructure for Big Data and Analytics is a next-generation platform with new processors, faster memory, and more storage options. It is designed, tested, and validated for enterprises to lower the cost of ownership and to scale from small to very large deployments as applications demand. With Cisco Application Centric Infrastructure (Cisco ACI™), it can scale to thousands of nodes. Cisco UCS delivers an optimal combination of high availability, performance, and flexibility while protecting your long-term investments.

Reference

- For more information about Cisco UCS, visit https://www.cisco.com/go/ucs.
- For more information about Cisco UCS big data solutions, visit https://www.cisco.com/go/bigdata.
- For more information about Cisco's big data validated designs, visit https://www.cisco.com/go/bigdata_design.
- For more information about Cisco UCS Integrated Infrastructure for Big Data and Analytics, visit https://blogs. cisco.com/datacenter/cpav5.