White paper

Modernizing Data Platforms for AI/ML and Gen AI

The benefits of migrating from Hadoop to Teradata Vantage

teradata.

Table of contents

Executive summary

Organizations today face a critical inflection point in their analytics journeys as they add AI/ML initiatives and generative AI (gen AI) capabilities to their existing traditional analytical data management workloads. Our analysis of a typical large enterprise environment with 100 TB of analytics data reveals substantial differences in operational efficiency, resource requirements, and associated costs across three platform options.

Existing Hadoop environments constrain business agility through their fragmented architecture. On-premises Hadoop deployments often require coordinating 10+ separate Hadoop components, each with its own development cycle and expertise needs. This complexity makes organizations reluctant to implement changes as systems grow, with modifications often causing cascading issues that increase technical debt. Cloud-based Hadoop options—such as Cloudera's Data Platform (CDP) offering—reduces some infrastructure challenges but maintains the same fundamental deployment complexity.

The staffing impact is substantial. On-premises Hadoop typically requires 21 to 28 full-time employees (FTEs) at an annual cost of $3.2 million to $4.2 million. Even cloud implementations need 13 to 18 FTEs costing $2.0 million to $2.7 million. By contrast, Teradata Vantage needs only three database administrators (DBAs), regardless of deployment model, saving $2.7 million to $3.6 million annually compared to on-premises Hadoop.

Technical resource requirement differences are equally dramatic. On-premises Hadoop requires at least 15 to 20 times more central processing unit (CPU) resources than Teradata Vantage for equivalent mixed data management workloads. Also, its combined issues with denormalization and storage replication demands significantly more physical storage for 100 TB of usable space reference environment. This inefficiency impacts sustainability, with on-premises deployments generating 1,500 to 3,000 metric tons of CO_2 annually—approximately 15 to 20 times more than Teradata Vantage.

While cloud-based Hadoop solutions such as Cloudera CDP offer some improvements to the areas of technical resources and storage replication, they fail to address the fundamental complexity of the Hadoop ecosystem. For organizations growing their analytical practices and embracing the promise of AI/ML and gen AI, Teradata Vantage provides a comprehensive solution through its unified architecture and consistent implementation across deployment models, enabling a strategic path toward a more efficient, future-ready analytics ecosystem.

The Hidden Cost Gap: Why Hadoop Requires More Resources Than Modern Alternatives

Platform Comparison	On-Premises Hadoop	Cloud-based Hadoop (Cloudera CDP)	Teradata Vantage
Platform Complexity	10+ separate Hadoop components with different upgrade cycles	10+ separate Hadoop components with varying cloud compatibility	Single integrated platform
Staffing Requirements	21 – 28 FTEs	13 – 18 FTEs	3 FTEs
Annual Personnel Cost	$3.2M – $4.2M	$2.0M – $2.7M	$420K – $540K
CPU Resources	15x – 20x	8x – 12x	1x (baseline)
Storage for 100TB	1,500 – 3,000 TB (1.5 – 3 PB)	500 – 1,000 TB (0.5 – 1 PB)	About 150 TB
Storage Efficiency	10x – 20x	3.3x – 6.6x	1x (baseline)
Energy Consumption	15x – 20x	5x – 7x	1x (baseline)

Introduction

Organizations today understand the importance and value of corporate-wide analytics. Both data warehouse and data lake/lakehouse architectures have evolved from nice-to-have to must-have components of corporate infrastructure. To remain competitive, businesses must be able to make critical strategic and tactical decisions based on as much data as possible, with AI/ML and generative AI (gen AI) capabilities becoming increasingly essential to this process.

Many enterprises have deployed Hadoop ecosystems over the past 10 to 15 years to support their analytics requirements. These organizations invested in Hadoop's promise of scalable, cost-effective processing for large volumes of diverse data. While these deployments initially may have met basic analytics needs, they now face a critical inflection point as business requirements evolve beyond what existing Hadoop environments can efficiently support.

The efficiency gaps between Hadoop and more modern platforms become increasingly pronounced as advanced data management workloads grow. Organizations find themselves constrained by Hadoop's inherent limitations—operational complexity, resource inefficiency, staffing challenges, and environmental impact. This new reality requires a strategic reevaluation of analytics infrastructure to ensure continued competitiveness and innovation.

For organizations embracing AI/ML and gen AI, the foundation of trusted data—seamlessly integrated and harmonized across the organization—is essential for success. Without this foundation of reliability, accuracy, and governance, investments in AI won't deliver their expected returns. Legacy Hadoop environments that were designed for an earlier era of analytics often struggle to provide this foundation, requiring organizations to chart a new course for their data infrastructure that can support current needs while adapting to future requirements.

Enterprise analytics evolution: A reference environment

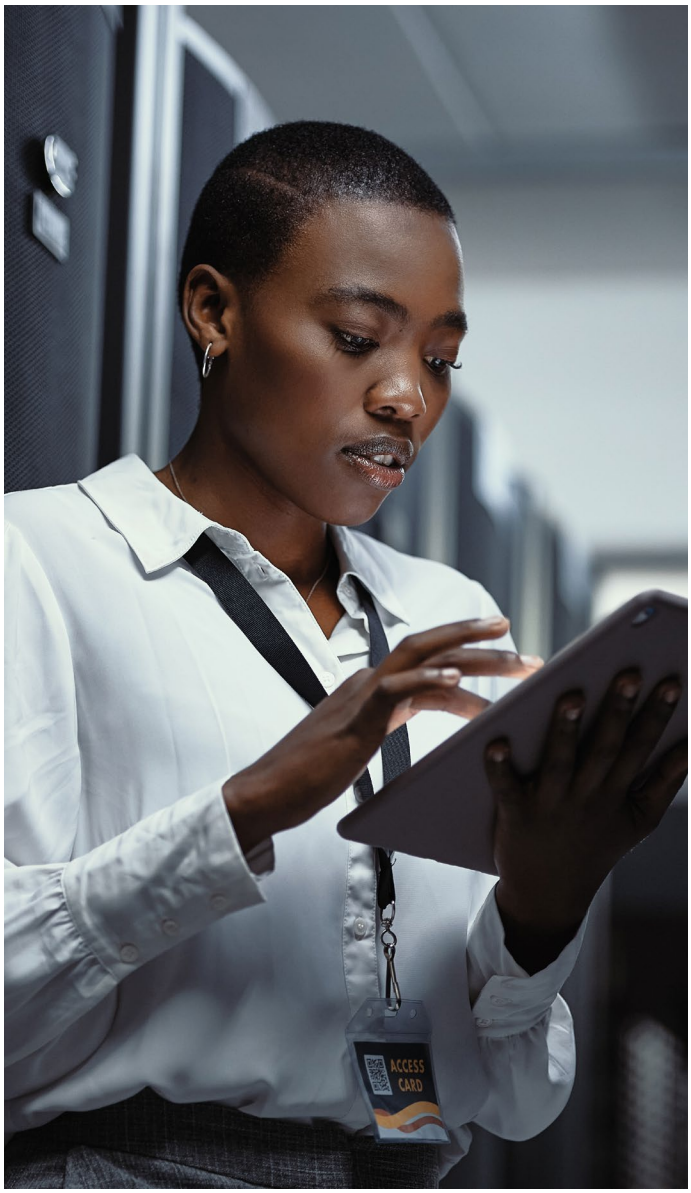
Our analysis examines a typical large enterprise with approximately 5,000 employees that has reached a critical inflection point in its analytics journey. With a current data environment of approximately 100 TB of data under storage representing concepts such as customer, product, and transaction, this environment supports hundreds of traditional analytics users and thousands of dashboard data consumers across its various business units. It's also nurturing a growing team of data scientists and ML engineers driving innovation.

Today's data management workloads reflect the organization's evolving priorities across three primary areas:

- Data transformation and preparation to integrate and clean organizational data
- Traditional SQL-based analytics supporting business reporting and dashboards
- Emerging data science practices utilizing AI/ML models for advanced analytics that enable more sophisticated insights

The analytics landscape is poised for significant transformation over the next 12 to 18 months as the organization expands into gen AI capabilities. Plans include implementing document summarization, semantic search, and internal digital assistant functionality for data analysis—innovations that will require substantial new infrastructure. This expansion demands specialized resources for vector databases and document storage.

These next-generation analytical workloads will place unprecedented demands on the organization's infrastructure. The way the organization addresses these evolving requirements will significantly impact not only its operational efficiency and costs, but ultimately its ability to deliver competitive advantage through data-driven analytics.



Hadoop and its challenges

Hadoop emerged 15 years ago as a solution for big data challenges, enabling organizations to store and process vast amounts of information across distributed computing clusters. Built as an open-source ecosystem, it allowed enterprises to leverage commodity hardware for data processing that was previously impossible or too expensive with traditional technologies.

Hadoop successfully addressed some data volume challenges of the past. However, it did little to solve the operational complexity of data management requirements. Today's analytical environments have fundamentally shifted into overdrive, with new demands to make data accessible to a wide range of business users across the enterprise. These transformative use cases include growing AI requirements—AI/ML, gen AI, and Bring Your Own Large Language Models (BYO-LLMs).

Strategic challenges with Hadoop environments

- **Operational efficiency:** Hadoop environments consist of large numbers of loosely associated open-source developed technology projects supported by the Apache Software Foundation. This complexity limits agility and makes organizations reluctant to implement changes, with modifications frequently causing cascading issues that increase technical debt.
- **Staffing burden:** Because of the considerable number of required technology components, specialized expertise is required to deploy a Hadoop ecosystem. This can create knowledge silos and operational risks as Hadoop skills become scarcer and/or more expensive.
- **Resource inefficiency:** Based on its commodity hardware heritage and distributed development history, Hadoop requires larger processing, memory, and storage footprints, requiring constant over-provisioning and frequent re-architecting as requirements evolve.
- **Energy consumption:** Excessive hardware footprints undermine sustainability commitments and increases operational costs through higher energy consumption.

Operational efficiency

Existing Hadoop environments significantly constrain business agility in several ways. Provisioning additional resources for new business programs, particularly AI/ML initiatives, typically can take weeks in on-premises Hadoop deployments. Quarterly maintenance windows often require 12 to 36 hours of downtime, directly impacting system availability and business operations.

As Hadoop environments grow in size and complexity, organizations become increasingly reluctant to implement necessary changes. Systems become harder to manage, tune, and navigate, making even locating relevant data challenging. Changes to the environment take progressively longer to implement and frequently result in unintended negative consequences, creating a cycle of increasing technical debt.

Components? Projects? Modules?

What's the right name for the pieces of the Hadoop ecosystem, such as Hive, Ranger, and Sentry?

Technically these are considered Apache projects within the Apache Software Foundation's development "process". Others might call these separate Hadoop technologies modules since they provide different functional business value to the Hadoop ecosystem.

Throughout this paper, we'll use the term Hadoop components to recognize the differences in nomenclature between technical and business contexts.

The complexity of the Hadoop ecosystem

Unlike integrated database platforms, Hadoop is fundamentally an ecosystem of separate open-source modules. For a mixed workload data management environment supporting data transformation, SQL analytics, and AI/ML capabilities, organizations must implement and maintain these numerous specialized Hadoop components, each requiring specific technical expertise—and staffing resources—to deploy, configure, tune, and troubleshoot.

The following is just an example of the types of components that need to be integrated to support our reference environment:

Hadoop Component	Primary Function
HDFS	Distributed file system for storage
YARN	Resource management and job scheduling
Hive/Impala	SQL processing and data warehousing
Spark	In-memory processing for batch/streaming
HBase	NoSQL database for random data access
Ranger	Security framework for authorization
Knox	Gateway service for REST APIs and UI
Oozie	Workflow scheduler for jobs
Kafka	Streaming data platform
Zookeeper	Coordination service for distributed systems

Over time, this fragmented architecture fosters organizational resistance to change—a "fear factor" where updates are delayed, technical debt accumulates, and outdated components remain in production. These operational inefficiencies directly impact business agility by extending time to market for new analytics capabilities, discouraging experimentation, and diverting resources toward ecosystem maintenance rather than business value creation. As requirements evolve, each new capability may require integrating yet another specialized component, further compounding these operational challenges.

Staffing

The fragmented architecture and operational complexities of Hadoop directly translate into significant staffing requirements. Each component in the ecosystem requires specialized expertise, creating a head-count-intensive support model that compounds the operational challenges already discussed.

From a staffing perspective, on-premises Hadoop typically requires 21 to 28 FTEs for comprehensive support, including system administrators, database administrators, Hadoop component support personnel for concepts such as security, and workload management. This staffing overhead translates to approximately $3.2 million to 4.2 million in annual personnel costs.

Personnel Requirements for Hadoop Environments (Based on U.S. market rates)

Role	Key Responsibilities	On-Premises Hadoop	Cloudera CDP on AWS
System Administrators	Hardware management, OS maintenance, capacity planning	4 – 5 FTEs	1 – 2 FTEs
DBAs	Query optimization, schema design, performance tuning	2 – 3 FTEs	2 – 3 FTEs
Hadoop Engineers	HDFS management, cluster tuning, MapReduce optimization	6 – 8 FTEs	3 – 4 FTEs
Data Engineers	ETL pipeline development, data transformation, integration	4 – 5 FTEs	3 – 4 FTEs
AI/ML Specialists	Model development, training, deployment, monitoring	3 – 4 FTEs	3 – 4 FTEs
Security	Security configuration, compliance, auditing	2 – 3 FTEs	1 FTE
Total FTEs		21 – 28 FTEs	13 – 18 FTEs
Annual Personnel Cost		$3.2M – $4.2M	$2.0M – $2.7M

Each component in the Hadoop ecosystem requires not just initial implementation expertise but ongoing operational support to ensure component-level performance, security, and compatibility. As organizations adopt more components to support diverse data management workloads, the staffing requirements will grow proportionally.

The challenge extends beyond just head count—these specialized roles are becoming increasingly difficult to fill as the market for Hadoop expertise continues to contract. Organizations often find themselves dependent on a limited pool of specialists whose knowledge may be limited to specific components rather than the ecosystem as a whole, creating risk through knowledge silos.

Technology environment challenges

Compute inefficiencies

Hadoop environments require substantially higher CPU resources compared to integrated analytics platforms due to their fragmented architecture. The need to support multiple independent components—such as core Hadoop with HDFS, MapReduce and YARN, Apache Hive for database or Apache Spark for in-memory processing—creates compute overhead through coordination and serialization processes between components. On-premises Hadoop deployments typically require at least 15 to 20 times the CPU resources of optimized analytics platforms to deliver complex mixed workloads.

This inefficiency stems from Hadoop's distributed processing model, uneconomical query optimization across components, and the fundamental overhead of data movement between separate processing engines. As organizations add AI/ML initiatives to their analytics environments, these compute inefficiencies become increasingly costly, often requiring significant hardware expansion.

Memory constraints

Memory management presents a critical challenge in Hadoop environments. The fragmented architecture lacks efficient memory-aware resource allocation, creating frequent bottlenecks for complex analytics. This limitation becomes particularly problematic for organizations considering expansion into data science and AI/ML workloads. These workloads typically require substantial memory for model training and inference. Deep learning models and large language models demand memory-intensive operations that Hadoop struggles to accommodate efficiently. Organizations typically over-provision memory to compensate for these limitations, further increasing infrastructure costs. When expanding AI capabilities, Hadoop environments often require complete re-architecting of memory allocation strategies and substantial hardware investments rather than simple capacity adjustments.

Joins vs. denormalization

Table joins and denormalization represent opposite approaches to handling related data for analytics.

Performant relational databases such as Teradata use table joins to combine data maintained in separate tables—such as customer, order, product. These table joins combine various aspects of two or more tables when the query is executed. The database only combines or "joins" the data when needed. This approach minimizes data redundancy, lowers storage requirements, and increases analytical flexibility.

Less sophisticated platforms such as Hadoop use denormalization rather than table joins. Denormalization pre-combines data from multiple tables into a single "wide" table. This approach avoids join operations, which are difficult—if not impossible—for Hadoop to perform efficiently. While this can improve query performance, it substantially increases both data preparation time in terms of building and rebuilding the denormalized tables and storage requirements through extensive data duplication.

Using denormalization with Hadoop often dramatically expands storage requirements by five to 10 times over the use of table joins with Teradata.

Storage constraints

On-premises Hadoop environments face substantial storage inefficiencies. To link different areas, such as customer, product, or region, existing Hadoop deployments typically require denormalizing analytical schemas to avoid costly table join operations between those datasets. This practice of denormalization greatly expands storage requirements—often by five to 10 times compared to a standard analytics environment, such as a data warehouse. This expansion is then multiplied by the requirements of Hadoop's HDFS architecture. HDFS requires that the data be replicated by a replication factor of three for data reliability and fault tolerance.

Consequently, a 100 TB analytics dataset can require 1.5 to 3 petabytes of physical storage in Hadoop to meet this requirement. Also, the tight coupling of compute and storage in on-premises Hadoop means storage expansions often require adding unnecessary compute capacity, creating cascading inefficiencies across the technology stack.

This contrasts sharply with traditional enterprise database systems, which typically need only about 1.5 times the analytical data footprint—150 TB for our 100 TB reference environment. This comes from the concept that enterprise database systems can perform numerous table joins that mixed analytical workloads currently require and avoid the costly practice of denormalization. This relatively low storage requirement includes database software and operations overhead, failover, and other performance requirements.

Energy consumption and carbon footprint

Energy consumption presents a substantial sustainability challenge for Hadoop environments. The extensive hardware footprint required by on-premises Hadoop's storage architecture and data processing profile directly translates to an enormous carbon footprint. Typical Hadoop deployments for a 100 TB environment would generate approximately 1,500 to 3,000 metric tons of CO_2 equivalent emissions annually. This is based on the requirements of "always on" disk storage and CPU processing power to maintain the environment.

This environmental impact—equivalent to the carbon sequestration of approximately 25,000 to 50,000 trees—creates significant concerns for organizations with sustainability commitments and ESG reporting requirements. Even during periods of low analytical demand, Hadoop clusters must maintain operation across all nodes to preserve data availability, resulting in continuous energy consumption regardless of actual processing needs.

Further environmental impact

The energy inefficiency of Hadoop environments has broader environmental implications beyond direct carbon emissions. The expanded hardware footprint demands proportionally more data center cooling capacity, which often consumes as much energy as the computing equipment itself, further compounding the environmental impact. Organizations face growing pressure from stakeholders, regulators, and customers to minimize their environmental footprint, making Hadoop's inefficient resource utilization increasingly problematic from a corporate responsibility perspective. Even cloud-based Hadoop deployments, while more energy-efficient than on-premises implementations, still generate substantially more carbon emissions than purpose-built analytics platforms due to their underlying architectural inefficiencies. This sustainability gap becomes more significant as organizations establish and commit to formal carbon reduction targets.

Teradata Vantage®: Addressing Hadoop's challenges

Teradata has built a four-decade-long heritage of creating trusted data foundations for mission-critical data management workloads, evolving from its data warehouse origins into a modern integrated analytics platform. Teradata Vantage fundamentally reimagines the analytics ecosystem by unifying data and analytical functions into a cohesive environment that eliminates the fragmentation inherent in Hadoop. This unified approach delivers dramatically better performance while requiring significantly less infrastructure and specialized expertise, resulting in significant cost advantages.

With its integrated ClearScape Analytics® capabilities, Vantage brings advanced AI/ML functionality directly into the platform, eliminating the need for separate specialized teams to support the infrastructure for this expanding—and critically important—practice. This creates a human-centric approach to analytics where business users and data scientists collaborate effectively using a common platform and shared data resources. The platform's integrated AI capabilities build organizational confidence in analytical outcomes and accelerate adoption and value creation across the enterprise.

Teradata's open and connected strategy extends to a customer's gen AI deployment strategies as well. By taking an open approach for modern AI infrastructure, Teradata supports three distinct design patterns:

- In-database execution leverages CPU processing for small language models, offering cost-effective solutions for task-specific needs
- In-platform deployment utilizes GPU infrastructure for medium-sized models, ideal for enhanced capabilities and regulatory use-cases requiring data privacy
- Model endpoint integration connects to external foundational LLMs like Google Gemini and OpenAI, enabling high-accuracy gen AI applications for conversational use-cases

All deployment options enable customers to deploy their chosen gen AI strategy without being constrained by the technology. This maintains Teradata's commitment to scalability and enterprise-grade performance.

Vantage's deployment flexibility—supporting on-premises, public cloud, and hybrid implementations with consistent functionality—allows organizations to implement their analytical data management workloads optimally based on business requirements, regulatory needs, and existing investments. This architectural consistency ensures the same principles of governance, security, and performance apply regardless of deployment model. For enterprises struggling with Hadoop's complexity and resource demands, Teradata Vantage offers a proven path forward that addresses each of Hadoop's strategic challenges while providing a foundation for future analytical innovation.

Teradata Vantage fundamentally reimagines the analytics ecosystem by unifying data and analytical functions into a cohesive environment that eliminates the fragmentation inherent in Hadoop.

Teradata's answers to Hadoop's strategic challenges

Operational efficiency

- **Hadoop challenge:** Loosely associated open-source components limit agility and increase technical debt through cascading changes.
- **Teradata solution:** Vantage provides a unified platform that eliminates cross-technology dependencies, reducing complexity and accelerating innovation. Vantage also embraces an open and connected strategy, seamlessly integrating with diverse data sources and technologies to enhance flexibility and ensure your data ecosystem remains agile and future-proof.

Staffing burden

- **Hadoop challenge:** Multiple components require specialized expertise, creating knowledge silos.
- **Teradata solution:** Vantage's integrated approach requires a smaller staffing head count compared to Hadoop, eliminating the need for specialized resources across multiple Hadoop components.

Resource inefficiency

- **Hadoop challenge:** Distributed architecture requires larger processing, memory, and storage footprints with constant over-provisioning of technical resources.
- **Teradata solution:** Vantage delivers superior analytical performance with vastly fewer compute/memory resources through its optimized query processing architecture, world class workload management, and considerably lower dedicated storage through its optimized architecture and efficient use of data storage.

Energy consumption

- **Hadoop challenge:** Excessive hardware footprints undermine sustainability commitments and increase energy costs.
- **Teradata solution:** Vantage's smaller infrastructure footprint means reduced carbon emissions and significantly lower power and cooling costs through greater computational efficiency.

Operational efficiency

Unlike Hadoop's fragmented ecosystem, Teradata Vantage provides a unified platform that eliminates the complexity of managing multiple components. To deliver mixed analytical workloads, Hadoop often requires coordinating the integration of 10 or more separate open-source developed components—each with different upgrade cycles and expertise demands. On the other hand, Vantage delivers a cohesive environment with consistent management interfaces and operational procedures. This architectural difference transforms the user experience, reducing complexity and allowing organizations to focus on analytics value rather than infrastructure maintenance.

	On-premises Hadoop	Cloud-based Hadoop	Teradata
Components required for support	10+ Hadoop components	10+ Hadoop components	A single integrated analytics platform

Vantage's unified architecture directly addresses the "fear factor" common in Hadoop environments. System updates apply to a single platform instead of requiring orchestration across multiple components, eliminating version compatibility issues that often prevent organizations from keeping their analytics environments current. By removing these inefficiencies, Teradata enables organizations to redirect resources from maintenance to innovation, accelerate analytics delivery, and maintain a modern platform that adapts to evolving business requirements without accumulating technical debt. Vantage also offers consistent implementation across deployment options—on-premises, cloud, or hybrid.

Staffing efficiency with Teradata Vantage

Vantage's unified architecture directly translates into dramatic staffing efficiencies. Unlike the fragmented Hadoop ecosystem that requires specialized expertise for each component, Teradata Vantage's integrated platform needs only 3 DBAs to support our complex mixed analytical workload 100 TB reference environment—regardless of whether the deployment is on premises or in the cloud. This consistent staffing model eliminates the need to develop different skill sets for different deployment models, further reducing operational complexity and training requirements.

The table below illustrates the stark contrast in personnel requirements across platforms:

Role	On-premises Hadoop	Cloudera CDP on AWS	Teradata Vantage
System Administrators (SA)	4 – 5 FTEs	1 – 2 FTEs	0 FTEs
Database Administrators (DBA)	2 – 3 FTEs	2 – 3 FTEs	3 FTEs
Core Hadoop Engineers for cluster and process management	6 – 8 FTEs	3 – 4 FTEs	N/A
Data transformation component specialists for Pig, Airflow, Oozie, etc.	4 – 5 FTEs	3 – 4 FTEs	0 FTEs
AI/ML component specialists for Mahout, MADlib, etc.	3 – 4 FTEs	3 – 4 FTEs	0 FTEs
Security components specialists for Ranger/Sentry, Knox, etc.	2 – 3 FTEs	1 FTE	0 FTEs
Total FTEs	21 – 28 FTEs	13 – 18 FTEs	3 FTEs
Annual Personnel Cost	$3.2M – $4.2M	$2.0M – $2.7M	$420K – $540K

This staffing reduction creates substantial annual cost savings—$2.8 million to $3.6 million compared to on-premises Hadoop and $1.6 million to $2.1 million compared to cloud-based Hadoop deployments like Cloudera CDP.

These savings stem from Teradata's integrated approach where system administration, data engineering, AI/ML capabilities (through ClearScape Analytics), and security functions are built into the platform, requiring no dedicated FTEs for infrastructure support and maintenance after initial installation.

Technology and environmental advantages

Superior compute efficiency

Teradata Vantage delivers significantly higher computational efficiency compared to Hadoop environments through its unified analytics architecture. By integrating analytical processing into a cohesive system rather than coordinating multiple disparate open-source driven technology components, Vantage eliminates the substantial coordination overhead inherent in Hadoop. The differences in CPU requirements across platforms are dramatic:

Platform	Relative CPU Requirements	Impact on Analytics Performance
On-premises Hadoop	15x – 20x	Higher latency, resource contention
Cloud-based Hadoop	8x – 12x	Improved but still highly inefficient
Teradata Vantage	1x (baseline)	Optimized performance, reduced processing overhead

This architectural advantage enables Vantage to support mixed analytical data management workloads with approximately 93% to 95% fewer CPU resources than on-premises Hadoop. The platform's sophisticated query optimization automatically determines the most efficient processing approach across workloads, eliminating the performance penalties typically seen when data moves between separate Hadoop processing engines. This efficiency becomes particularly critical for AI/ML workloads, where the computational challenges that Hadoop environments struggle to overcome efficiently. Vantage's integrated ClearScape Analytics capabilities leverage the platform's optimized compute resources without requiring additional processing layers or the redundant storage that dramatically increases Hadoop's computational footprint.

Optimized memory utilization
Teradata Vantage's massively parallel processing (MPP) architecture fundamentally transforms memory utilization compared to Hadoop's similar, and yet highly fragmented, approach. Vantage's world-class workload management system intelligently coordinates analytical workloads based on business priorities and resource requirements. This approach eliminates the memory bottlenecks common in Hadoop environments by ensuring critical workloads receive appropriate resources without manual intervention. The platform's sophisticated query optimizer maximizes memory efficiency by determining optimal execution paths. For AI/ML workloads, Vantage leverages these same optimizations to efficiently handle memory-intensive operations better than Hadoop environments, enabling organizations to expand analytical capabilities.

Storage efficiency
Teradata Vantage dramatically reduces storage requirements through its data processing capabilities and its overall efficient use of storage resources. The platform's storage efficiency creates substantial differences in physical storage needs across deployment options:

Platform	Physical Storage Required for 100 TB Environment	Storage Efficiency Ratio	Primary Storage Mechanism
On-Premises Hadoop	1,500 – 3,000 TB (1.5 – 3PB)	10x – 20x	Required denormalization, HDFS with 3x replication
Cloud-Based Hadoop	500 – 1,000 TB (0.5 – 1PB)	3.3x – 6.6x	Required denormalization
Teradata Vantage	150 TB	1x	Advanced table join efficiency with integrated redundancy

This efficiency stems from Vantage's ability to efficiently process complex table joins without requiring data denormalization, along with its optimized storage architecture. The dramatically smaller storage footprint reduces not only direct storage costs but also associated infrastructure for power, cooling, and management. For organizations with substantial data growth projections, Vantage's storage efficiency becomes increasingly valuable as it allows analytical capabilities to scale without the associated growth in storage infrastructure that Hadoop environments require. This difference becomes even more pronounced when considering the data expansion needs of modern AI/ML workloads, where multiple versions of very large datasets may need to be maintained.

Reduced environmental impact
Teradata Vantage's technological efficiency translates directly into environmental benefits. With its dramatically smaller infrastructure footprint, Vantage environments generate approximately 150 to 200 metric tons of CO_2 equivalent emissions annually for our 100 TB reference deployment—roughly 90% to 95% less than comparable Hadoop environments. This substantial carbon reduction helps organizations meet increasingly stringent ESG goals and sustainability commitments. The platform's efficient resource utilization is significantly better than the continuous energy consumption pattern common in always-on, on premises Hadoop clusters. Vantage's hybrid cloud capabilities further enhance sustainability by allowing data management workloads to run in the most energy-efficient environment, whether on premises or in cloud regions powered by renewable energy.

Seamless hybrid cloud implementation
Teradata Vantage provides consistent implementation across on-premises, public cloud, and hybrid configurations through its powerful QueryGrid® technology. QueryGrid serves as the backbone for seamless data and workload integration across environments, allowing organizations to deploy analytical workloads optimally while maintaining identical functionality and operational procedures. This technology enables bidirectional data movement and query federation across Vantage environments regardless of deployment model, eliminating the complex data synchronization challenges—including significant network bandwidth and cloud egress charges—that plague hybrid Hadoop deployments. Organizations can leverage QueryGrid to maintain a unified analytics ecosystem that spans deployment strategies while ensuring consistent performance, security, and governance.

Choosing the right path forward

As organizations reach a critical inflection point in their analytics journey, decisions about platform modernization must balance immediate needs with long-term strategic goals. While on-premises Hadoop environments have served enterprises well for traditional analytics workloads, they present significant challenges for the future—particularly as analytics requirements evolve toward AI/ML and gen AI capabilities. The resource inefficiencies of Hadoop ecosystems—requiring at least 15 to 20 times more CPU, significantly more memory and storage, and 10 to 15 times greater energy consumption—create technical limitations and financial burdens that grow increasingly problematic as the demands of AI/ML workloads expand.

Cloud-based Hadoop implementations such as Cloudera Data Platform (CDP) offer only partial relief from these challenges. While they reduce some physical infrastructure management through object storage and CSP managed services, they still require substantial specialized expertise across multiple Hadoop components. Organizations adopting cloud Hadoop still need 13 to 18 FTEs with specialized Hadoop knowledge, cloud-specific skills, and expertise in identity management, networking, and cost optimization. This approach fails to address the fundamental complexity and inefficiency of the Hadoop ecosystem itself.

Teradata Vantage provides a comprehensive solution to these challenges through its integrated, unified architecture. With staffing requirements reduced to just 3 DBAs for our reference 100 TB environment, Vantage eliminates the need for systems administrators, specialized Hadoop component experts to support data engineering, AI/ML, and security infrastructure. Its built-in ClearScape Analytics capabilities deliver advanced AI functionality without the need for additional technical resources, while its consistent implementation across on-premises, cloud, and hybrid environments allows organizations to deploy analytics where they make the most business sense.

We recommend a strategic approach to modernization focused on business value. Through this process, organizations can identify high-value analytics, declutter their technology estates of underutilized assets, and create streamlined paths to more efficient, sustainable, and future-ready analytics ecosystems with Teradata Vantage.

Next steps

Building on a strategic assessment of your analytics environment, Teradata offers a structured migration path from an existing Hadoop ecosystem deployment to Vantage through our comprehensive Hadoop migration program:

1. **Hadoop migration platform assessment service:** We begin with a detailed evaluation of your existing Hadoop environment, examining system configuration, architecture, workloads, and data utilization patterns. This assessment provides a complete picture of migration scope, potential risks, and realistic timelines while identifying opportunities to eliminate unused or low-value components.
2. **Hadoop migration planning service:** Based on assessment findings, we develop a comprehensive migration strategy aligned with your business objectives. This phase creates a targeted architecture and technology roadmap for both the migration process and target platform, ensuring your highest-value analytics are prioritized for early migration.
3. **Hadoop migration implementation service:** The execution phase moves selected components from your Hadoop environment to Teradata Vantage, including data, schemas, applications, and pipelines. Throughout implementation, we enhance data quality, access patterns, and lineage tracking to deliver improved analytics capabilities—not just a platform change.

Also, Teradata's QueryGrid technology can serve as a bridge between your existing Hadoop deployment and a new Vantage implementation, enabling seamless operations during the transition period. This allows for staged migration without disruption to business operations, providing immediate benefits from Vantage while maintaining access to Hadoop-based assets until migration is complete.

Appendix

Energy consumption and sustainability impact background

Power consumption information is derived by combining industry standard data center energy consumption metrics with comparative efficiency analyses between the different platforms. The base calculation starts with estimating power consumption for a typical 100 TB Hadoop cluster, considering:

- Server requirements, including storage and processing requirements
- Typical power draw per server
- Data center overhead, including cooling and power distribution
- Average PUE (power usage effectiveness) for enterprise data centers

For the emissions calculations:

1. **On-premises Hadoop (1,500 – 3,000 metric tons CO_2e):** Based on estimated annual power consumption of approximately six to 12 million kWh, using average U.S. grid carbon intensity
2. **Teradata Vantage (150 to 200 metric tons CO_2e):** Derived from the efficiency advantage established earlier, applying the same grid carbon intensity factors for on-premises deployments as well as cloud service providers' (CSPs) more efficient data centers when deployed in a cloud-environment
3. **Cloud-based Hadoop (500 – 1,000 metric tons CO_2e):** Positioned as a middle ground based on:
   - CSP efficient data centers (PUE typically 1.1 to 1.2 versus 1.6 to 2.0 for enterprise data centers)
   - More modern hardware with better performance per watt
   - Still requiring more infrastructure than Teradata due to fundamental architectural differences

The tree sequestration equivalency (25,000 to 50,000 trees) was calculated using Arbor Day Foundation estimates that a mature tree absorbs approximately 48 pounds of CO_2 annually.[1]

It should be noted that these power consumption and sequestration numbers represent reasonable approximations based on commonly available information rather than specific measurements from any single deployment. They're intended to illustrate the relative scale of environmental impact between the platforms. Specific power consumption and sequestration numbers can only be determined on a case-by-case basis.

Staffing requirements and cost comparison

Staffing FTEs are blended from various sources on enterprise Hadoop deployments across various industries. This blended information shows that organizations typically need 20 to 30 FTEs to maintain a production-grade Hadoop environment to support a 100 TB environment with the stated mixed analytical data management workloads and supporting an organization of 5,000 employees.

The annual personnel costs were calculated using median salary data from multiple sources—such as Indeed, Glassdoor, and Salary.com—including industry compensation surveys, job market analysis platforms, and enterprise IT staffing reports. The 2025 estimated annual fully loaded costs (including benefits, taxes, and overhead) for each role in the U.S. market are:

- System administrators: $120K – $150K
- Database administrators: $140K – $180K
- Core Hadoop (HDFS, MapReduce, YARN) engineers: $150K – $190K
- Transformation (Pig, Airflow, Oozie) component support: $130K – $170K
- AI/ML component (Mahout, MADlib) specialists: $160K – $220K
- Security component (Ranger/Sentry, Knox) engineers: $130K – $160K

It should be noted that actual staffing requirements and costs may vary based on your organization's size, geographic location, complexity of workloads, existing skill sets, and specific implementation details. These examples are intended to illustrate the relative scale of staffing and costs.

1. "The Value of Trees," Arbor Day Foundation, https://www.arborday.org/value.