



Data modernization
solutions paper
www.cirata.com

Your legacy data stack *is holding you back.*

How migrating from Hadoop and other legacy on-premises data lakes to modern cloud platforms unlocks AI, analytics, and competitive advantage.



The data architecture of the past *cannot power the enterprise of today.*

The data landscape has fundamentally shifted. What Hadoop and on-premises data lakes represented a decade ago, a breakthrough in scalable, affordable big data storage, has become, for most organizations, a significant liability. These systems were built for a different era of data processing: batch-oriented, compute-coupled, and optimized for stability over agility. They were never designed to feed the real-time analytics, machine learning models, and AI applications that define competitive advantage in 2026.

The consequences are visible across industries. Data science teams wait hours for jobs that should take minutes. Engineering resources are consumed by

maintaining complex, brittle infrastructure rather than building new capabilities. AI initiatives stall not because the models are insufficient, but because the data they need cannot be delivered at the speed, scale, or quality required. Meanwhile, cloud-native competitors, unencumbered by legacy infrastructure, move faster and extract more value from the same raw data.

Data modernization is not a technology upgrade. It is a strategic repositioning, the deliberate decision to replace infrastructure that constrains what is possible with infrastructure that enables it. Cirata exists to make that transition fast, safe, and continuous, regardless of the scale or complexity of the data estate being migrated.

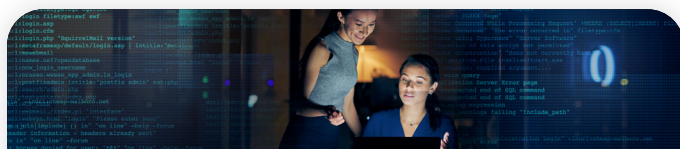


"Over **70%** of enterprises have already begun migrating workloads from legacy systems to modern platforms, driven by the need for AI-ready, scalable infrastructure."

IDC Enterprise Data Migration Survey

Three forces that are *reshaping the risk landscape.*

The economics, velocity, and character of digital disruption have shifted materially. Three overlapping forces define the environment organizations must now plan to recover from.



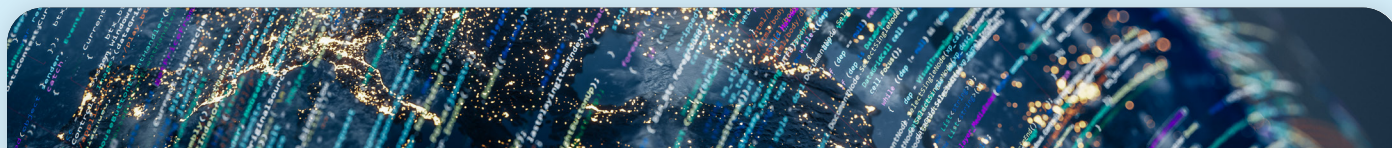
Cyber incidents have become operational outages

Ransomware and destructive attacks are no longer primarily about data exfiltration. They are engineered to interrupt core processes, degrade environments, and force extended downtime. The Verizon 2025 Data Breach Investigations Report identifies ransomware as a dominant feature of breaches, tightly linked to system intrusion events. The target is operational continuity and recovery is the battlefield.



Infrastructure concentration creates systemic exposure

Cloud and edge providers deliver extraordinary scale and reliability – until failure modes cascade across regions, control planes, identity layers, and dependent services. Architectures described as ‘multi-region’ or ‘highly available’ often share hidden dependencies in DNS, IAM, CI/CD pipelines, or administrative tooling. When these shared components fail, redundancy assumptions collapse. The Cloudflare outage of December 2025 illustrated how services can become unreachable even while underlying compute and storage remain fully intact.



Complexity amplifies recovery cost

IBM’s Cost of a Data Breach Report (2025) estimates the global average cost of a breach at \$4.44 million and \$10.22 million in the United States. More consequentially, breaches spanning multiple environments (public cloud, private cloud, and on-premises) take an average of 276 days to identify and contain, compared to 207 days for single-environment incidents, and cost \$5.05 million on average. The complexity of distributed data estates is not just an architectural challenge, it is a direct multiplier on recovery cost and time.

Legacy data infrastructure *has a compounding cost.*

The instinct to delay modernization is understandable. Migration carries risk, complexity, and cost, all visible, all immediate. What is less visible, but far larger, is the cost of not modernizing: the accumulated drag of systems that require constant maintenance, cannot scale efficiently, and block access to the capabilities that define modern competitive performance.

70% of enterprises have begun migrating workloads from legacy systems to modern data lake or lakehouse platforms.

50%+ reduction in time-to-value reported by organizations moving from Hadoop to cloud-native platforms.

78% of enterprises now require ML infrastructure, making AI-ready data platforms a baseline requirement, not a differentiator.

Source: IDC Enterprise Data Migration Survey;
Gartner Infrastructure & Operations Survey

The financial case for modernization is reinforced at every level. Operational cost savings from eliminating on-premises hardware refresh cycles, reducing specialist Hadoop administration overhead, and moving to elastic cloud pricing are well-documented. But the more significant gains are strategic: the

ability to access AI and machine learning platforms that simply do not operate on legacy architectures; the freedom to choose analytics tools without being locked to a single vendor's ecosystem; and the capacity to move data to where it creates value, rather than where it was first stored.

Four structural limitations Hadoop *was never designed to solve.*

Legacy Hadoop environments were engineered for a specific purpose: cost-effective, distributed storage and batch processing of large datasets. They performed that function well. But the demands placed on enterprise data infrastructure have changed entirely, and the architectural assumptions baked into Hadoop cannot be retrofitted to meet them.

01.

Tightly coupled storage and compute cannot scale independently

Hadoop's architecture binds HDFS storage and MapReduce compute together. Scaling either one requires scaling both, creating significant inefficiency and cost. Modern workloads, particularly AI training jobs and real-time analytics, require the ability to spin compute up and down independently of storage. This is structurally impossible in a traditional Hadoop cluster, and it is the single most common driver of over-provisioning, wasted spend, and performance ceilings.

02.

AI and machine learning cannot run effectively on HDFS data

Modern AI and ML platforms are built to operate on cloud-native object storage with open table formats such as Apache Iceberg. They do not integrate natively with HDFS. Organizations attempting to run AI initiatives on legacy Hadoop environments face constant data extraction overhead, format conversion bottlenecks, and architectural mismatches that slow model development and reduce output quality. There is no practical path to scalable AI without first modernizing the data layer.

03.

Operational overhead grows as the technology ecosystem moves on

Hadoop administration requires specialized expertise, in HDFS management, YARN configuration, Hive metastore maintenance, and cluster operations, that is increasingly scarce and expensive. As the vendor ecosystem has moved toward cloud-native platforms, tool compatibility has narrowed, and the cost of maintaining Hadoop currency has risen. Engineering teams that should be building analytics and AI capabilities are consumed by infrastructure maintenance. The opportunity cost compounds every quarter.

04.

Data governance and compliance become harder, not easier

Modern data governance requirements, including fine-grained access control, full data lineage, ACID transactions, and schema evolution, are either difficult or impossible to implement cleanly on legacy Hadoop stacks. Open table formats like Apache Iceberg, which underpin modern lakehouse architectures, provide these capabilities natively. As regulatory requirements tighten and AI governance obligations grow, the governance gap between legacy and modern data infrastructure becomes a direct compliance risk, not just an operational inconvenience.

A cloud-native foundation built for *next level AI, analytics, and scale.*

Data modernization is not simply about replacing one storage system with another. It is about moving to an architecture that was designed from the ground up for the capabilities that matter: elastic scale, real-time access, open interoperability, and native AI/ML integration. Understanding what that architecture looks like is essential to planning the journey.



Cloud-native object storage

Replacing HDFS with cloud object storage, decouples storage from compute, enables elastic scaling, and dramatically reduces the operational overhead of managing distributed file systems. Storage costs drop; access speeds improve; and the data becomes available to the full ecosystem of cloud-native analytics and AI tools.



Open table formats: Apache Iceberg

Apache Iceberg has become the de facto standard for open, interoperable table formats in modern data lakehouses. It provides ACID transactions, schema evolution, time travel, and partition pruning, capabilities that Hive cannot deliver. Critically, Iceberg is engine-agnostic, which eliminates vendor lock-in at the data layer.



AI and analytics platform Integration

The leading AI and analytics platforms are all built for cloud-native data. Once data is modernized onto cloud object storage in open formats, these platforms become immediately accessible, enabling the full spectrum of use cases from real-time dashboards to large language model fine-tuning.



Elastic compute and governance

Modern cloud architectures allow compute resources to scale dynamically with workload demand, eliminating the over-provisioning that defines legacy Hadoop clusters. Paired with unified governance organizations gain fine-grained access control, full data lineage, and compliance tooling that meets the requirements of GDPR, CCPA, DORA, and emerging AI governance frameworks.

Recoverability engineered *into the data estate.*

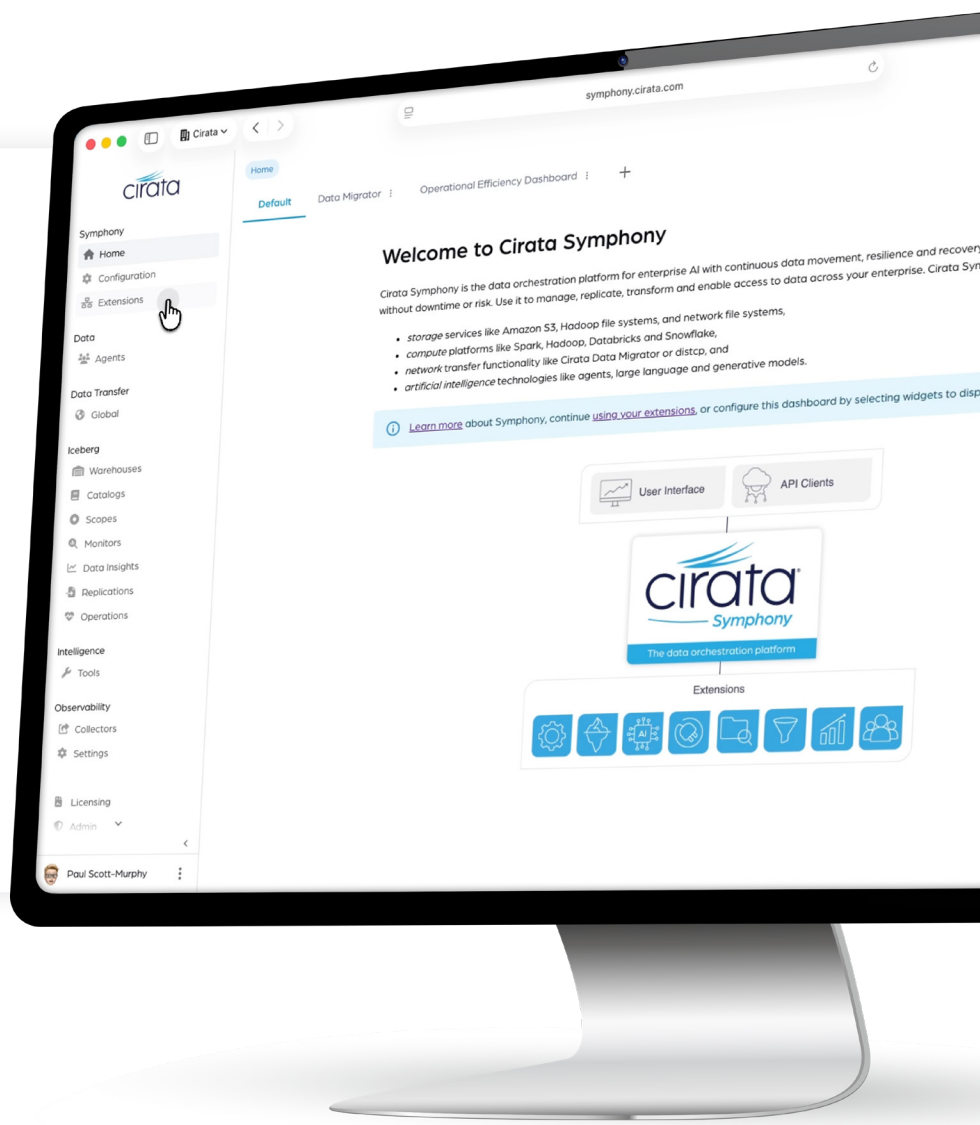
Cirata Symphony is an intelligent data orchestration platform built for the scale and complexity that enterprise AI demands. Cirata Symphony moves and manages data across any environment, on-premises to cloud, cloud to cloud, or across regions securely, continuously, and without disruption to live systems.

Unlike legacy migration tools that require downtime, complex custom scripting, or repeated manual reconciliation, Symphony automates the entire process. It captures changes as they happen, ensuring that the data available to AI and analytics teams is always current and complete.

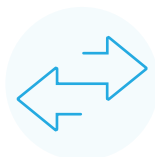
Freeing your teams to do what they do best

When data engineers spend their time maintaining fragile integration scripts, they are not building models.

Cirata Symphony eliminates the operational burden of data movement, redirecting your most valuable technical talent toward the work that actually drives business value.



What Cirata Symphony delivers.



Real-time orchestration

Continuous movement of data and metadata. Move datasets of any size from on-premises data lakes to any cloud environment without disrupting production workloads.



Zero disruption

Changes to source data are captured and replicated in real time without risk, code rewrites, or downtime, so models are never trained on stale information.



Freedom and flexibility

Native support for Apache Iceberg and other open table formats ensures data remains portable and future-proof.



Visibility and control

Full oversight of usage, cost, and security with automation that eliminates the need for hand-crafted scripts, freeing engineering teams to focus on building.



Enterprise trust

Single connective layer for mission-critical, regulated environments

What becomes possible when *legacy constraints are removed.*

Data modernization is not an end state, it is an enabler. The value of moving from legacy Hadoop to a modern cloud data stack is realized not in the migration itself, but in what the migration makes possible. The following capabilities are consistently unlocked once organizations complete the transition to cloud-native, open-format data infrastructure.

Real-time analytics and business intelligence

Legacy Hadoop environments process data in batch cycles, hourly, daily, or longer. Modern cloud data platforms enable streaming ingestion and query execution against live data, compressing decision cycles from days to minutes. Organizations gain the ability to act on current information rather than yesterday's snapshot: dynamic pricing, real-time fraud detection, live inventory management, and instant customer personalization all become architecturally feasible.

Machine learning at scale

ML model training requires iterative access to large, clean datasets at high speed. On modern cloud platforms, data scientists can provision elastic compute clusters, access data directly from cloud object storage in open formats, and run training jobs in parallel without competing for cluster resources with production workloads. The result: faster model iteration cycles, better model quality, and shorter paths from experimentation to production deployment.

AI and retrieval-augmented generation

Large language models and generative AI applications require continuous access to fresh, high-quality organizational data. This is architecturally incompatible with Hadoop. Once data is modernized onto cloud-native infrastructure with continuous orchestration, organizations can implement RAG pipelines that allow AI systems to query live organizational data, powering intelligent assistants, automated reporting, and AI-driven customer experiences that were previously out of reach.

Freedom of platform choice

One of the most significant benefits of modernization is architectural freedom. Data stored in open formats like Apache Iceberg is accessible to any engine without duplication or vendor-specific conversion. Organizations can adopt the best analytics tool for each use case, switch providers as the market evolves, and avoid the lock-in that has historically made technology transitions so costly.

Unified governance and compliance

Modern data platforms provide the governance capabilities that AI adoption and regulatory compliance now require: fine-grained access control, complete data lineage, ACID transactions for data integrity, and audit trails that satisfy GDPR, CCPA, and sector-specific requirements. Organizations moving to open lakehouse architectures can implement governance policies once and enforce them across every tool that accesses the data.

Westpac begins to exploit their *digital and hybrid multi-cloud strategy*

Westpac is Australia's first bank and oldest company, one of four major banking organizations in Australia and one of the largest banks in New Zealand. The bank is halfway through their 5 year program to modernize its technology environment and expand use of cloud-based systems. To achieve this Westpac, have partnerships with both Microsoft Azure and Accenture.

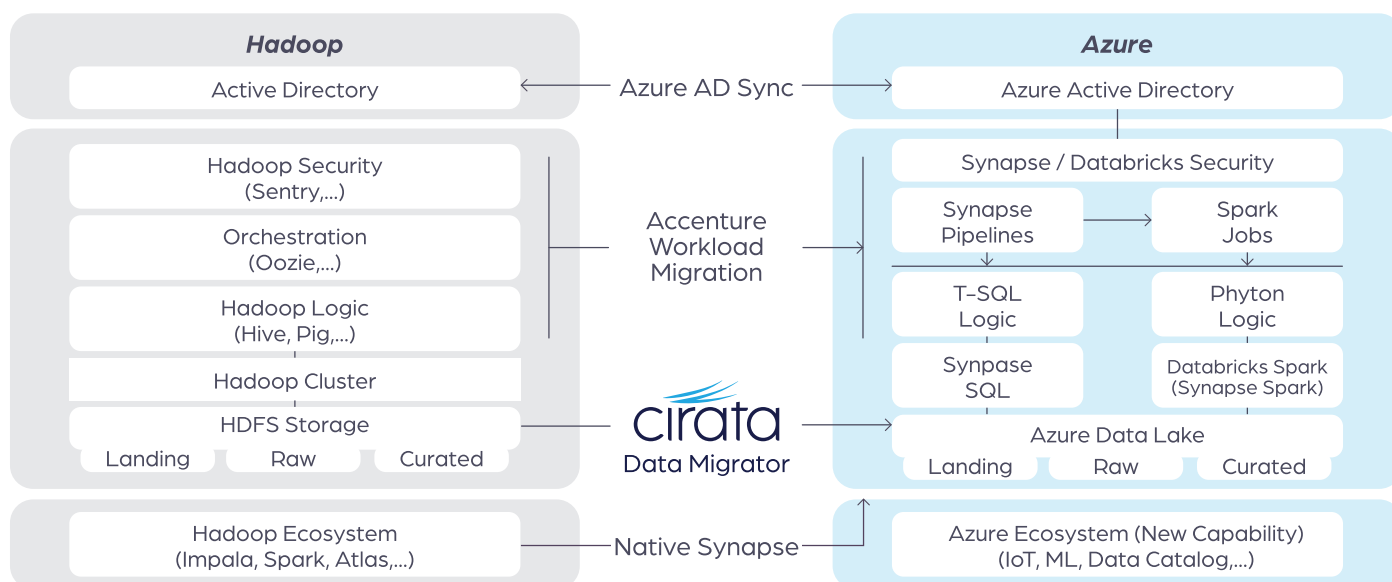
Westpac is focused on

-  Reducing technology duplication
-  Making it easy to discover customer insights
-  Providing unique customer interactions

Cirata data migrator was used to successfully move their Hadoop data into the Azure Data Lake, and once migrated also move this data to other domains within the Azure infrastructure. The data migration was a live migration in that any updates that occurred to the Hadoop cluster during data migration were also captured as part of the migration process, ensuring that the cloud version of the data was in sync with the original.

Results

- Automated the data transfer from the on-premise Hadoop cluster to Microsoft Azure.
- Optimized data transfer performance and network bandwidth usage.
- Provided the data to enable upstream applications access to the data in the cloud. This required data movement between ADLS domains.
- 2 PB of data moved to the cloud.





Begin your data modernization journey *with Cirata.*

Data modernization at enterprise scale is a significant undertaking, but it does not have to be a disruptive one. The organizations that complete migrations fastest and

most successfully are those that approach the journey in phases, with the right tooling and a clear view of what they are migrating toward and why.

Contact us to learn more about Cirata Symphony and how we can help you *move your data from legacy to cloud-native, at any scale, without disruption.*

cirata.com/data-orchestration/solutions/data-modernization