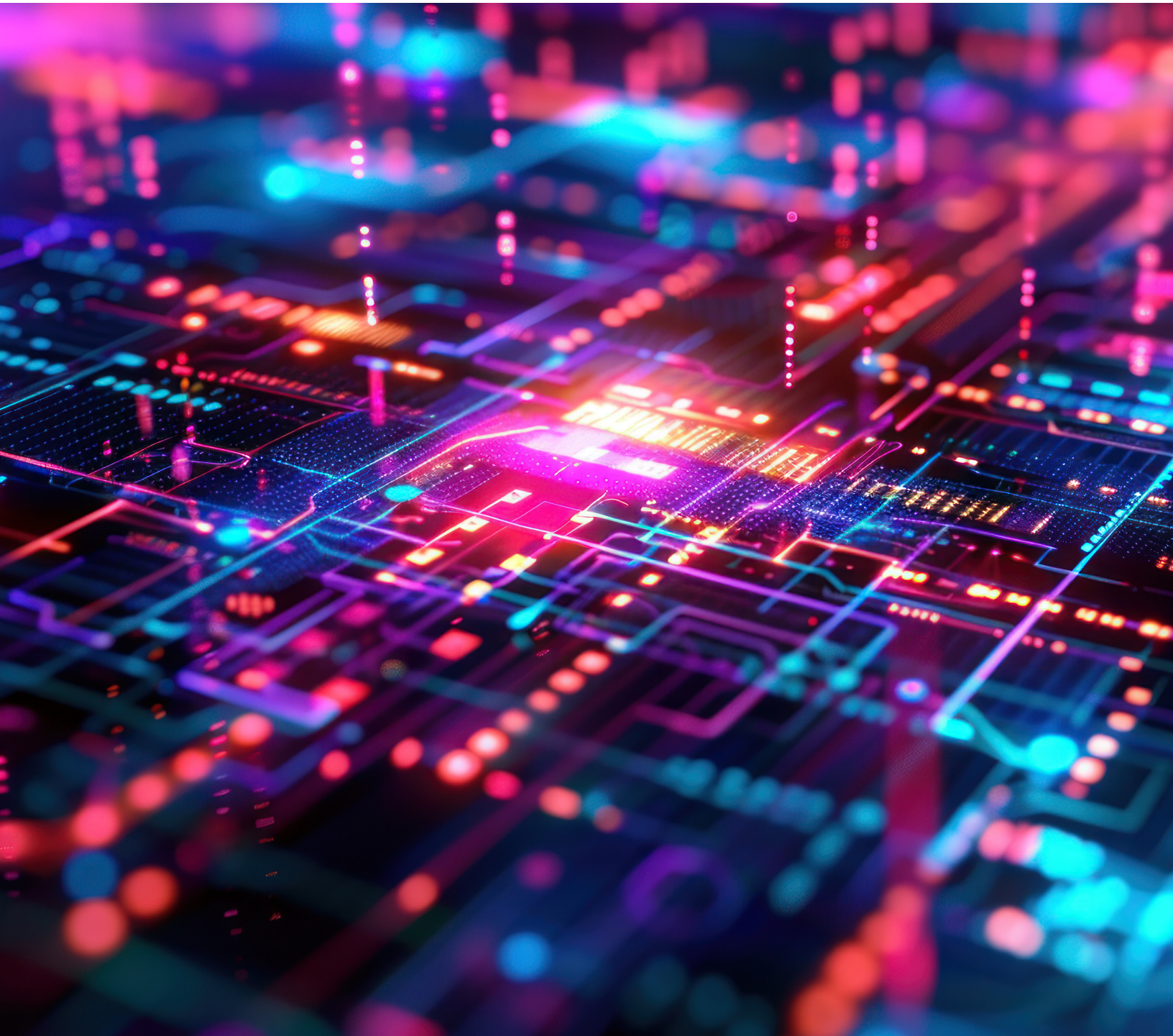
riverbed

WHITE PAPER

# Eliminating AI Data Bottlenecks at Enterprise Scale

## Riverbed Data Express: Data Movement at the Speed of Innovation



# Contents



# Executive Summary

In today's AI-powered enterprise, data is the fuel — and its movement is the engine. Yet, as organizations scale AI initiatives across cloud, edge and data center environments, they face a growing crisis: data movement bottlenecks that delay insights, waste expensive GPU resources and erode competitive advantage.

This white paper explores the challenges in data movement to feed AI applications, and the strategic imperative of agile, secure, and cost-effective data movement approach. It then introduces **Riverbed's Data Express Service** that dramatically accelerates data transfer, reduces costs, and enhances observability across multi-cloud environments.

# The Strategic Imperative: Why Data Movement Matters More Than Ever

AI is hungry — and data is the feast.
It's increasingly important to move data to AI models rather than the other way around due to the growing scale and complexity of modern AI systems. Large models, especially those used in enterprise and cloud environments, require specialized infrastructure that isn't easily portable.

AI models, especially large-scale ones like foundation models or LLMs depend on GPUs, TPUs, or custom accelerators, along with optimized software stacks and orchestration tools. These GPUs are concentrated in specialized data centers because they require high-performance infrastructure that's optimized for intensive computational workloads like AI training and inference. These data centers are equipped with advanced cooling systems, high-speed networking, and scalable power supplies to support the massive energy and heat output of GPU clusters.

AI workloads, especially those involving large language models and real-time analytics, require **terabytes to petabytes of data** to be moved daily. Industries like financial services, healthcare, media, and manufacturing are generating unprecedented volumes of data — and the challenge is no longer just storing it but **moving it to where AI models live.**

> Moving data to AI is now the norm — because GPUs and AI infrastructure are rarely co-located with enterprise data.

AI models rely on various types of data, including structured data (like databases and spreadsheets), unstructured data (such as text, images, audio, and video), and semi-structured data (like logs and JSON files). This data is typically created or stored in enterprise systems, cloud platforms, edge devices, IoT sensors, and user-facing applications. For example, customer interaction data may reside in CRM systems, operational logs in observability platforms, and image data in cloud storage or mobile devices.

The result is the data needed to feed AI models running in these specialized data centers is likely to reside somewhere else, either in other cloud environments, private data centers or edge locations. Thus in order to feed the AI model, the data must be moved from where it resides to where the AI model runs. This presents four distinct challenges.

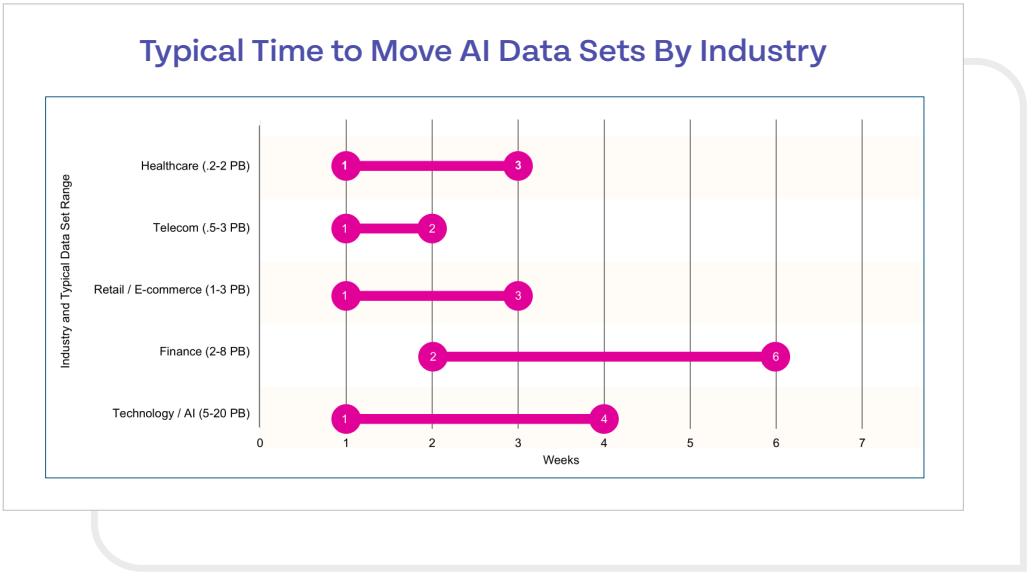
# Top Four Challenges in AI Data Movement

## Time

Moving the petabytes of data needed to feed AI applications present significant time-related challenges. Even with high-speed connections, transferring this volume can take days or weeks, creating delays that are unacceptable for time-sensitive AI use cases. For example, over a perfectly-tuned 10 Gbps connection, it would take roughly 9 days to move 1 PB of data continuously.

However, different industries need to regularly move different amounts of data, and their different infrastructure investment can greatly impact how fast they can move 1 PB of data.
(See "Typical Time to Move AI Data Sets By Industry" graphic.)

Before data can be moved, it often needs to be cleaned, transformed, and validated. These preprocessing steps add significant time overhead, especially when dealing with unstructured or inconsistent data sources.




**Sources:** Microsoft, Initus, moonswitch.com

Distance, inadequate network infrastructure and protocol overhead can slow down data movement, especially when data must be moved from remote geographic locations to optimized data centers running AI models. These bottlenecks can delay model training, inference, and decision-making processes.

AI models often require frequent retraining with updated data as stale data reduces model accuracy and business value. Frequent retraining of AI models with updated data compounds the issue, requiring repeated cycles of data movement and processing. Each cycle involves not just moving data, but also reprocessing and revalidating it, which adds to the time burden and can slow down innovation or responsiveness. Altogether, these challenges make it difficult to maintain data freshness and responsiveness in AI systems, especially at enterprise scale.

There are two related dimensions to time in moving data to AI: scale and complexity. AI data movement often means petabytes of data, and coordinating the movement of large datasets across environments requires complex scheduling and orchestration, which can introduce delays due to dependencies, resource availability, and error handling. Let's look at each.

## Scale

The problem scale creates in moving data for AI is multifaceted and increasingly severe as data volumes grow. IDC's 2024 InfoBrief Activating Enterprise Data with AI and Analytics highlights that enterprises are increasingly combining internal structured data with external structured and unstructured data to train and operate AI models. This includes text, images, video, and telemetry data—often amounting to petabytes per month for large organizations.

While neither IDC nor Gartner provides a direct monthly petabyte figure, their infrastructure and spending forecasts strongly suggest that large enterprises are moving multiple petabytes of data monthly to support AI workloads—especially in cloud and hybrid environments.

The sheer size of modern datasets—often in terabytes or petabytes—consumes significant bandwidth and compute resources, which can strain infrastructure and reduce overall system efficiency. This can lead to latency issues, especially in time-sensitive applications like real-time analytics or AIOps. These are systemic bottlenecks that no amount of GPU scaling can fix.

### Common Industry-Specific Data Set Sizes and Use Cases

| Industry | Daily Data Volume | Key AI Use Cases |
|---|---|---|
| Financial Services | 1–100+ TB | Fraud detection, trading, credit scoring |
| Healthcare | 1–50 TB | Imaging, diagnostics, patient monitoring |
| Retail & E-commerce | 10–100+ TB | Recommendations, pricing, sentiment |
| Manufacturing | 1–10 TB | Predictive maintenance, robotics |
| Media & Entertainment | 10–100+ TB | Streaming, targeting, moderation |

### Not One and Done

Keeping data fresh for AI models is essential to maintain accuracy, but the process of constantly ingesting and transferring data introduces delays that can undermine time-sensitive applications.

This places a heavy burden on resources, both human and infrastructure, consuming staff time, bandwidth, compute, and storage resources—especially when dealing with real-time or high-volume data. Managing these transfers adds operational complexity, requiring robust orchestration and governance, often requiring new, dedicated teams to properly perform these functions, as the risk to the business of not doing it correctly is too great to do otherwise.

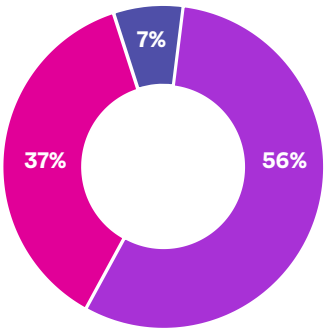
 

## Complexity

Beyond the sheer scale required to move data to AI applications, there is increased complexity. As data becomes more distributed across cloud, edge and on-premise environments, consolidating it for AI processing introduces security, privacy and compliance risks, particularly in regulated industries.

This fragmentation introduces challenges in data integration, latency, governance, and cost. Each environment may use different formats, APIs, security protocols and access controls, making it difficult to unify data for AI processing. Synchronizing data across these silos requires sophisticated orchestration and often leads to inconsistencies or delays, which can degrade model performance or lead to inaccurate insights.

### Where Enterprise Data Resides




**56%** of all enterprise data, structured and unstructured, is stored across multiple public clouds.

Additionally, transferring data between environments raises security and compliance risks, as data may cross jurisdictional boundaries or violate privacy regulations. The number of different and sometimes competing regulations creates a complex matrix of what can be moved when and to where. Just keeping up with the ever-changing compliance regulations of different jurisdictions is challenge onto itself.

Moving data from edge locations to centralized AI applications is challenging due to several layers of complexity. Edge locations often generate high volumes of real-time data that need to be processed quickly. However, these locations typically have limited bandwidth, compute power, and storage, making it difficult to transfer large datasets efficiently to centralized AI models.

Additionally, edge data is often heterogeneous and decentralized, coming from different formats, protocols, and systems. This requires extensive data normalization and integration before it can be used effectively by AI. The same security, compliance and latency challenges also apply to moving data from edge to AI applications.

## Cost

Perhaps biggest, most important cost is the opportunity cost in delaying delivery of business results with AI because of slow data movement. New revenue and business models relying on AI results are extremely time-sensitive and cannot reach their intended audience until data is delivered to the AI model and the results are then produced.

When data isn't delivered to models in a timely manner, it can lead to outdated insights, slower decision-making, and missed chances to respond to emerging trends or threats. For example, in AIOps or cybersecurity, delayed data can mean slower detection of anomalies or breaches, increasing risk and potential damage. In customer-facing applications, stale data can reduce personalization accuracy, leading to lower engagement or conversion rates.

These delays also impact on innovation cycles. AI models that rely on fresh data for retraining or fine-tuning may fall behind, causing enterprises to miss windows for product optimization, market responsiveness, or operational efficiency. Any delay in feeding the model risks losing first-mover advantage to competitors who deploy fresher models faster. Ultimately, the time lost in data movement translates into lost competitive advantage, reduced agility, and diminished ROI from AI investments.

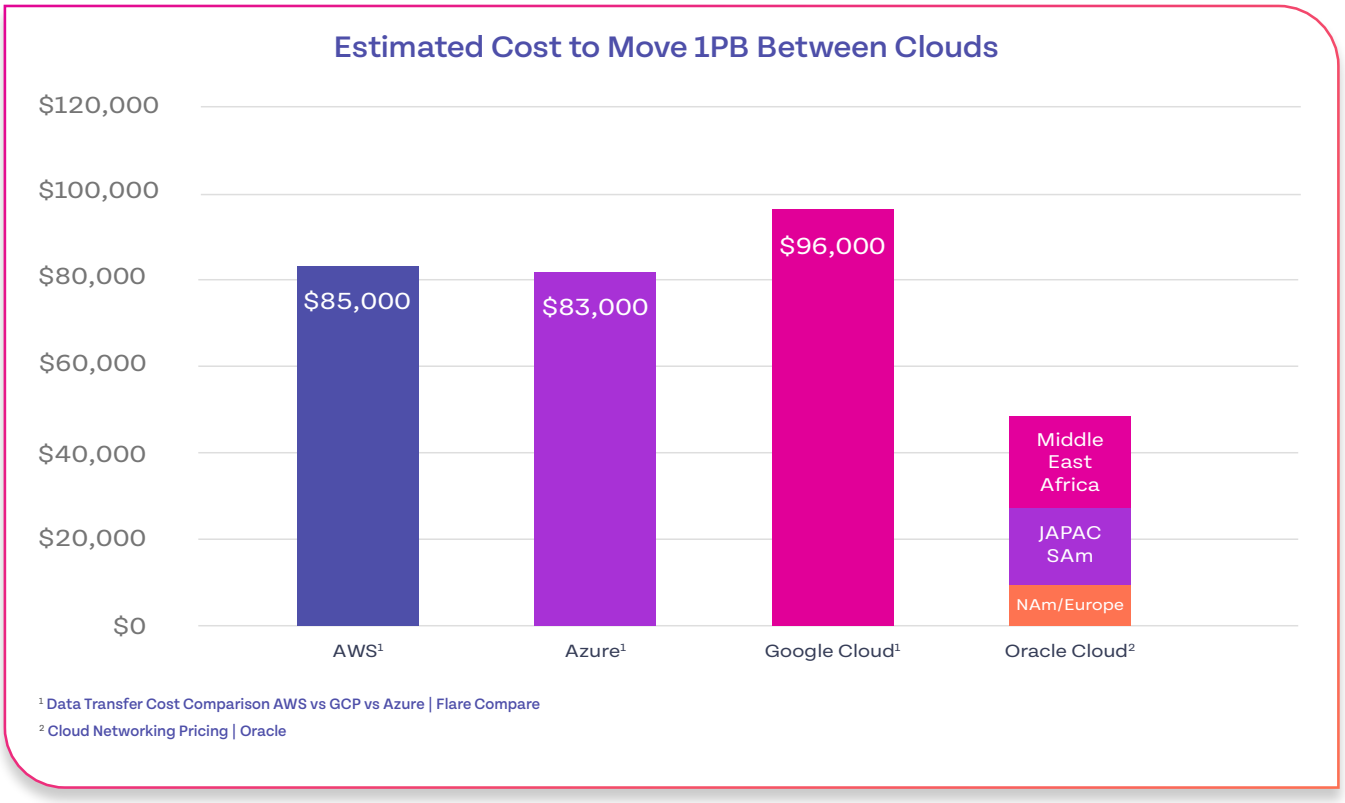
 

## Hard Costs, Too

Beyond lost opportunity costs, there are actual real costs in moving data. One of the most significant is cloud egress fees, which are charged when data is transferred out of cloud environments—between different cloud providers or different regions within the same cloud provider, or to on-premises systems. These fees can be substantial when dealing with petabytes of data.

It can easily cost $80,000 in egress charges to move 1 PB of data out of a cloud provider, although each of the four major hyperscalers have different rates depending on various factors including where the data resides, so actual costs do vary. (See "Estimated Cost to Move 1PB Between Clouds" graphic).

Beyond egress fees, moving data out of a cloud provider incurs several other hard and hidden costs that can significantly impact enterprise budgets, These may include network transfer and bandwidth costs, especially for high-throughput or cross-region transfers; storage tiering and duplication to prepare data for transfer; compute costs for data preparation which consume compute resources and can drive up cloud bills—particularly when done at scale, various licensing, API access fees, orchestration tools and monitoring expenses and of course labor costs.




Since data feeding AI needs to be constantly updated, these costs can be incurred multiple times every month.

Finally, there is a time-related data movement opportunity cost for the GPUs used in AI applications. Because demand for GPU capacity currently outstrips GPUs availability, cloud providers lease GPU capacity, holding it in reserve for customers to use. Organizations pay substantial amounts for this GPU capacity whether it is used or not. Thus, any delay in timely delivery of data to these GPUs has both real as well as opportunity costs. (See DataCrunch GPU Pricing Comparison for full details and pricing breakdowns.)

# Solving the Top Four Challenges in AI Data Movement

## Riverbed Data Express: SaaS-Based Simplicity

Sometimes, enterprises just want to move data. They want the easy button to move massive amounts of data across their multi-cloud infrastructure.

Riverbed Data Express Service enables enterprises to radically accelerate the movement of massive datasets now required to train and deploy AI models at scale—reaching data transfer speeds up to 10 times faster than current industry solutions —improving a customer's time to value and lowering costs. With the new Riverbed Data Express Service, what once took months can now be completed in days—giving organizations the speed and security of data delivery now required to prosper in the AI era.

## Fast: Move Petabytes Per Day—Literally

Riverbed Data Express Service is engineered for speed, offering transfer rates from 5TB to 40TB per hour—that's up to 1PB per day. With support for concurrent transfers and continuous sync, you can scale to meet even the most demanding AI or cloud migration workloads. Compared to alternative available solutions, Data Express Service delivers nearly 10x performance at peak.

## Easy: SaaS Simplicity and Better Economics

Forget complex configurations or hidden costs. Data Express Service is a fully managed SaaS platform—no VMs to deploy or manage, no tuning required. The power of Riverbed SteelHead is under the covers, ready to move petabytes of data. You pay only for what you move, with a consumption-based pricing model that aligns with your cloud economics. For many data types, the service can reduce cloud egress costs up to 90%, reducing overall cost of data migration.

Setup takes minutes via a simple user interface and integration is seamless via REST APIs. It works across Oracle Cloud Infrastructure today, with planned services for multi-cloud (AWS, OCI, Azure, Google Cloud Platform) and data center connectivity. Watch the demo

## Secure: Built for Enterprise-Grade Protection

Security is non-negotiable. Data Express Service is multi-tenant by design, with RBAC, SSO/IAM, and tenant isolation at both the control and data layers. Data in transit is protected with post-quantum cryptography (PQC) secure tunnels.

You can even deploy agents inside your own cloud or data center fabric to apply your existing security policies.

## Insightful: Real-Time Visibility and Governance

From day one, you get deep visibility into job status, throughput, file counts, and history—all through a unified portal. Over time, Riverbed is adding outcome-oriented insights like time saved, top data movers, and trend analysis.

Advanced analytics and alerting integrate with your workflows, giving you the observability and governance you need to stay in control.

## Use Cases That Matter to CIOs

Riverbed Data Express Service is purpose-built for the modern CIO agenda:

- **AI/ML Training and Inference:** Feed GPU clusters faster to avoid idle compute.
- **Cloud-to-Cloud Transfers:** Support multi-cloud strategies with ease.
- **Data Center to Cloud Migrations:** Modernize apps and analytics.
- **Disaster Recovery:** Replicate data securely and quickly.
- **Edge-to-Core Consolidation:** Centralize IoT and remote data.

Whether you're in financial services, healthcare, media, or manufacturing, Data Express Service adapts to your enterprise needs. You can subscribe via Riverbed sales or partners, with options for pay-as-you-go or enterprise subscriptions. Training and onboarding are available through Riverbed Academy.

The Riverbed Data Express Service is currently in limited availability, with general availability coming soon. If you're a CIO looking to future-proof your data infrastructure, now is the time to engage.

# Data Movement As a Strategic Priority

Data movement no longer has to be cumbersome, backend task—it can be a strategic enabler. Riverbed Data Express Service gives you the speed, simplicity, security, and insights needed to lead in an AI data-driven world.

Ready to accelerate your AI journey? Contact us for a Riverbed Data Express demo.

riverbed

### About Riverbed

Riverbed, the leader in AI observability, helps organizations optimize their users' experiences by leveraging AI automation for the prevention, identification, and resolution of IT issues. With over 20 years of experience in data collection and AI and machine learning, Riverbed's open and AI-powered observability platform and solutions optimize digital experiences and greatly improve IT efficiency. Riverbed also offers industry-leading Acceleration solutions that provide fast, agile, secure acceleration of any app, over any network, to users anywhere. Together with our thousands of market-leading customers globally – including 95% of the FORTUNE 100 – we are empowering next-generation digital experiences. Learn more at riverbed.com.