Blue Orange

DATA INFRASTRUCTURE / MLOPS

SECTOR INTELLIGENCE REPORT

AI Data Optimization in Data Infrastructure & MLOps

Building the Scalable Foundation for AI-Driven Value Creation

Prepared for Operating Partners of Private Equity Firms

Blue Orange Digital | March 2026

March 2026

Table of Contents

Executive Summary

In December 2025, Databricks closed a $5 billion funding round at a $134 billion valuation, 2.3 times Snowflake's public market capitalization. The round included $2 billion in debt capacity, a structure that signals IPO preparation rather than typical growth-stage fundraising. Databricks' AI products alone generate $1.4 billion in annualized revenue, growing at 65%, within a business producing $4.8 billion total with positive free cash flow. The message to the market was clear: data infrastructure is not a cost center. It is the asset class that determines whether enterprise AI investments produce returns or become write-offs.

That distinction matters urgently in 2026. Enterprise AI spending reached $644 billion in 2025, but only 14% of CFOs report measurable ROI. Seventy-two percent of that investment is estimated to be destroying value through waste and poor measurement. Sixty percent of AI projects are abandoned due to insufficient data quality. Forty-two percent of companies scrapped most of their AI initiatives in 2025, up sharply from 17% the prior year. The pattern is consistent across industries: the bottleneck to AI value creation is not model capability, which has advanced dramatically. It is the data infrastructure layer that feeds, governs, monitors, and operationalizes those models.

This whitepaper examines the data infrastructure and MLOps landscape through the lens of private equity value creation. It maps the $660 to $690 billion in hyperscaler capital expenditure that is reshaping the demand environment, the platform consolidation dynamics that are creating acquisition opportunities, and the specific infrastructure investments that separate the 14% of organizations achieving AI ROI from the 86% that are not. Blue Orange Digital's AI Data Optimization Framework provides the structured methodology for diagnosing data readiness, prioritizing infrastructure investments, and sequencing deployments that compound into measurable EBITDA enhancement.

The $690 Billion Catalyst: Hyperscaler Infrastructure Spending

The scale of capital flowing into AI infrastructure in 2026 is historically unprecedented. Combined hyperscaler capital expenditure targets for the year range from $660 to $690 billion, a 36% increase over 2025. Amazon is planning $200 billion. Alphabet is targeting $175 to $185 billion. Meta has committed $115 to $135 billion. Microsoft is spending $120 billion or more. Oracle rounds out the group at $50 billion. Goldman Sachs projects cumulative hyperscaler capex of $1.15 trillion between 2025 and 2027, a 2.4 times increase over the $477 billion spent during 2022 to 2024.

These numbers would be remarkable in any context. What makes them alarming is the capital intensity they represent. Hyperscaler capex is now consuming 45 to 57% of revenue, against a historical norm of 10 to 15%. The top five hyperscalers are consuming 100% of operating cash flows for infrastructure investment, versus a ten-year average of 40%. For the first time, hyperscalers hold more debt than cash, with $121 billion in bonds issued in 2025 alone and projected debt issuance of $1.5 trillion between 2026 and 2028 to fund continued buildout.

The Power Bottleneck: The Constraint Nobody Expected

The primary constraint on AI infrastructure deployment in 2026 is not GPU availability. It is electricity. Microsoft CEO Satya Nadella publicly acknowledged that GPUs are sitting idle in inventory due to lack of power, with $80 billion in unfulfilled Azure orders attributable to electricity constraints rather than compute capacity. Power transformer lead times have stretched to 128 weeks, against an industry standard of 12 to 16 weeks. The International Energy Agency projects that global data center electricity consumption will double to 945 terawatt-hours by 2030.

For PE operating partners, the hyperscaler spending boom creates a dual dynamic. On the demand side, it is driving unprecedented investment in data infrastructure tools, platforms, and services across every enterprise that touches cloud computing. On the supply side, power constraints are creating new categories of value in power-efficient computing, edge infrastructure, and optimization technologies that reduce the compute footprint required for AI workloads. Portfolio companies positioned at either end of this dynamic will benefit disproportionately.

Data Quality: The $644 Billion Problem

The single most important statistic in enterprise AI is this: 60% of AI projects are abandoned due to insufficient data quality. Not model accuracy problems. Not deployment complexity. Not talent gaps. Data quality. The implication is that the majority of AI spending, across a $644 billion enterprise investment base, is being deployed against infrastructure that cannot support the workloads it is expected to power.

The failure cascade is predictable. An enterprise invests in a machine learning platform, hires data scientists, selects use cases with strong theoretical ROI, and then discovers that the data required to train and serve those models is fragmented across incompatible systems, riddled with inconsistencies, missing critical fields, and governed by policies that are either nonexistent or unenforceable. Forty-two percent of companies scrapped most of their AI initiatives in 2025, up from 17% in 2024, a trajectory that reflects not growing skepticism about AI but growing recognition that AI without data infrastructure is an expensive exercise in disappointment.

The Governance Gap: 62% Cite It as the Primary Impediment

Sixty-two percent of organizations identify data governance as their greatest impediment to AI advancement. The correlation between governance maturity and AI outcomes is stark: organizations that solve governance challenges deploy AI three times faster with 60% higher success rates. Mature governance correlates with 40% higher analytics ROI, 24.1% revenue improvement, and 25.4% cost savings. The financial cost of not investing in governance is equally measurable: more than 25% of organizations lose over $5 million annually due to poor data quality, and 7% report losses exceeding $25 million.

For PE operating partners, this data creates a clear value creation pathway. Data governance is not a technology project. It is an operating discipline that, when implemented systematically, produces predictable, compounding returns. The 40% ROI improvement from mature governance is achievable within 12 to 24 months and applies across every AI initiative the portfolio company subsequently deploys. In a landscape where 86% of enterprises cannot confidently measure AI ROI, the ability to implement governance that reliably produces measurable returns is a genuine competitive advantage.

Platform Consolidation: The Architecture Decisions That Determine Value

The data infrastructure market is consolidating around two architectural paradigms, and the choice between them has material implications for PE portfolio company strategy and exit valuations.

The Lakehouse Convergence

Databricks and Snowflake have converged on the data lakehouse architecture from opposite directions, and their current positioning illustrates the market dynamics. Databricks, built on Delta Lake with ACID transactions, schema evolution, time travel, and cost-effective cloud object storage, trades at a 25 times revenue multiple on its private valuation, reflecting the market's premium for its AI-native positioning and 55% growth rate. Snowflake, which has added Apache Iceberg table support to its proprietary micro-partitioned storage, trades at 15 times revenue publicly, a discount that reflects both its mature market position and the perception that its architectural origins in data warehousing carry more migration risk than Databricks' lakehouse-native approach.

The critical development is the standardization around open table formats. Both Delta Lake and Apache Iceberg are now widely supported across platforms, reducing vendor lock-in and enabling organizations to build architectures that work across clouds. For PE portfolio companies, open table format adoption is a strategic priority: it reduces switching costs, improves negotiating leverage with vendors, and positions the company for exit with a flexible infrastructure that acquirers can integrate without costly migration projects.

The MLOps Platform Market: $4.38 Billion and Consolidating

The MLOps market is projected to reach $4.38 billion in 2026, growing at 39.8% CAGR, but the more significant dynamic is consolidation. Databricks' acquisition of Tecton, the feature store leader, signals a strategy of absorbing standalone MLOps capabilities into the lakehouse platform. Feature stores, model registries, vector databases, and observability tools are converging into unified platforms rather than remaining independent product categories. MLflow 3.0, released in 2025, added first-class generative AI support including prompt tracing, LLM judge capabilities, and experiment metadata for GenAI workflows.

The ML CI/CD pipeline has also standardized. Production-grade MLOps now includes unit tests for data transformations, holdout evaluation for model performance validation, fairness checks and bias detection, LLM-specific tests for factuality and toxicity, and promotion gates requiring manual approval for high-risk models. The shift from experimental ML to governed, production ML mirrors the broader enterprise software maturation cycle, and PE operators who recognize this pattern can drive value by implementing production-grade MLOps discipline in portfolio companies still running ad hoc model development processes.

Vector Databases: Becoming Infrastructure, Not Product

The vector database market has rapidly matured from a standalone product category into an embedded infrastructure component. Pinecone leads in production environments with P99 latency of approximately 47 milliseconds at one billion vectors, compared to Weaviate's 123 milliseconds. Weaviate's advantage is native hybrid search, combining vector, keyword, and metadata queries in a single operation. Chroma serves the developer-first prototyping market but lacks enterprise multi-tenant capability.

The strategic trend is bundling: Databricks has integrated LanceDB, Snowflake has added vector search capabilities, and the standalone vector database market is being absorbed into broader data platforms. For PE operating partners, this means vector database capability is a checklist item for portfolio company data infrastructure, not a standalone investment thesis. The value sits in how vector search integrates with the broader RAG and agent infrastructure, not in the vector database itself.

Real-Time Data Processing: The AI Inference Backbone

The release of Apache Flink 2.2.0 in December 2025 marked an inflection point for real-time AI infrastructure. The release added ML_PREDICT for large language model inference directly in streaming pipelines and VECTOR_SEARCH for real-time vector similarity search in streaming context, with seamless integration into OpenAI, Anthropic, and Databricks AI ecosystems. The significance is architectural: for the first time, organizations can run AI inference as a continuous streaming operation rather than a batch process, eliminating the latency between data collection and model prediction.

Kafka and Flink together form the enterprise AI backbone for organizations operating at scale. Streaming agents, event-processing services that maintain state and trigger decisions, connect data streams to external AI model APIs for real-time anomaly detection, data enrichment, and automated decision-making. OpenAI uses Kafka and Flink for its generative AI data pipelines. Financial services firms run real-time fraud detection on streaming infrastructure. E-commerce platforms power real-time personalization through the same architecture.

The Streaming-First Lakehouse

The architectural convergence emerging in 2026 is the streaming-first lakehouse, where ingestion, processing, and query serving operate as continuous rather than batch operations. This architecture eliminates the traditional ETL lag between data arrival and availability for AI inference, enabling real-time feature computation that feeds directly into production models. For PE portfolio companies with latency-sensitive use cases, including fraud detection, dynamic pricing, personalization, and operational alerting, streaming-first architecture is becoming the baseline competitive requirement.

AI Agent Infrastructure: The Production Frontier

AI agent infrastructure has consolidated around three frameworks in early 2026. LangGraph 1.0, released in 2025, is the production leader for complex stateful workflows requiring deterministic execution and auditability, with a graph-based architecture where nodes represent functions and edges define execution flow. CrewAI, with 44,600 GitHub stars, leads in speed-to-prototype for role-based multi-agent orchestration, with deep MCP integration that automates tool discovery and connection lifecycle management. Microsoft's Agent Framework, reaching general availability in Q1 2026 as the successor to AutoGen, targets enterprise conversational agents with compliance and governance requirements.

The production adoption data is significant: 52% of executives report AI agents in production deployment, and 74% of those are achieving ROI within the first year. Compare that to the 14% of CFOs who can measure general AI ROI. The gap suggests that agent infrastructure, when paired with proper data infrastructure, produces dramatically higher success rates than undirected AI investment. For PE operating partners, this means agent deployment is the highest-confidence path to AI ROI, provided the data infrastructure foundation is in place.

Production Deployments: Case Studies with Measurable ROI

06

Sharp Business Systems: The Infrastructure-First Lesson

Sharp Business Systems, a global technology provider, deployed AI-powered sales intelligence with account-fit scores, intent signals, and organizational change detection. The initial deployment failed. Adoption was poor despite the AI capability itself functioning correctly. The root cause was data infrastructure: the systems feeding the AI tool could not produce real-time signals with sufficient accuracy or freshness to drive sales behavior change.

Sharp paused the AI deployment and spent six months rebuilding its data infrastructure, integrating real-time business intelligence tracking across organizational changes, technology installations, and buying signals from multiple sources. When they relaunched, adoption reached 80%. The lesson is direct: the AI model was never the bottleneck. The data layer was. For PE operating partners evaluating AI readiness, this case study argues that data infrastructure assessment should precede AI investment, not follow it.

Fortune 500 BI Impact: 5x Revenue Growth, 2.5x Valuation

ZoomInfo's analysis of Fortune 500 companies leveraging advanced business intelligence platforms found measurable performance differentials that PE investors should note: companies with mature BI infrastructure achieved 5 times the revenue growth, 89% higher profit margins, and 2.5 times the valuation multiple of their industry peers. The correlation between data infrastructure maturity and enterprise value creation is not theoretical. It is observable in public market data across the largest companies in the world.

Platform ROI: MuleSoft at 445%, Informatica at 335%

Forrester's Total Economic Impact study of MuleSoft's Anypoint Platform documented 445% ROI with $7.8 million in total benefits per implementation, driven by API-led connectivity that accelerated time-to-market, reduced integration costs, and improved data asset reusability. Nucleus Research's analysis of Informatica Cloud across five industries found 335% ROI over three years. These returns reflect the compounding nature of data infrastructure investment: once the integration and governance layer is built, every subsequent AI initiative benefits from it.

Company / Platform	Investment	Key Metric	ROI / Impact
Sharp Business Systems	Data infra rebuild (6 mo)	80% AI tool adoption	Failed to succeeded
Fortune 500 (ZoomInfo)	Advanced BI platforms	5x revenue growth	2.5x valuation premium
MuleSoft Anypoint	API-led connectivity	$7.8M benefits/impl.	445% ROI
Informatica Cloud	Data integration	5 industries studied	335% ROI / 3 years
AI Agent deployments	Agent infrastructure	52% in production	74% achieving ROI Y1

The EBITDA Expansion Playbook for Data Infrastructure

Data infrastructure investment follows a compounding logic that differs from other AI deployments. Each layer of infrastructure, including governance, integration, real-time processing, and ML operations, amplifies the return on every subsequent AI initiative. The portfolio company that builds a strong data foundation captures not just the direct value of governance improvement but the accelerated deployment and higher success rates of every AI project that follows.

The Three-Tier Deployment Sequence

Tier	Initiative	EBITDA Impact	Timeline	Success Rate
1	Data governance framework	40% analytics ROI lift	0-6 months	3x faster AI deployment
1	Data quality remediation	Prevent $5-25M losses	0-6 months	60% higher success
2	Open table format migration	Reduce vendor costs 20-30%	6-18 months	Multi-cloud flexibility
2	Real-time data pipelines	Enable latency-sensitive AI	6-18 months	Production AI backbone
3	AI agent infrastructure	24.1% revenue uplift	18-36 months	74% ROI in Y1
3	Production MLOps discipline	25.4% cost savings	18-36 months	Governed, auditable AI

The Valuation Impact

The exit multiple math reflects the compounding nature of data infrastructure investment. Mid-market data infrastructure companies trade at 8 to 12 times revenue, while Databricks commands 25 times and Snowflake 15 times. A PE-backed portfolio company that enters at 8 to 10 times EBITDA and implements systematic data governance, migrates to open table formats, deploys real-time pipelines, and achieves production AI with measurable ROI can realistically target 10 to 13 times EBITDA at exit. The 20 to 30% multiple uplift is driven by the same factors that differentiate the ZoomInfo Fortune 500 cohort: demonstrable correlation between data infrastructure maturity and financial performance.

Blue Orange Digital's AI Data Optimization Framework

Blue Orange Digital has built its practice around the insight that data infrastructure is the determining factor in AI ROI. The AI Data Optimization Framework provides a structured methodology for assessing data maturity, identifying the specific infrastructure gaps that are blocking AI value creation, and sequencing investments in the order that produces the fastest compounding returns.

The Composite Priority Score

Every infrastructure initiative is evaluated using a composite score that balances potential EBITDA impact against implementation feasibility:

Score = ((EBITDA Low + EBITDA High) / 2) x Portfolio Multiplier / (Data Readiness x Implementation Complexity x (Time to Value / 12))

The framework's insight is that data infrastructure initiatives have uniquely high Portfolio Multipliers because the same governance framework, integration architecture, or streaming pipeline can be replicated across every portfolio company. A PE fund with ten portfolio companies that implements a standardized data governance methodology captures ten times the return on the initial design investment.

The Data Maturity Assessment

The framework begins with a diagnostic that maps each portfolio company against five dimensions: data quality and completeness, governance and compliance, integration and accessibility, real-time processing capability, and ML operations maturity. The assessment produces a quantified readiness score for each potential AI initiative, enabling operating partners to distinguish between initiatives that can deploy immediately on existing infrastructure and those that require foundational investment before they can generate returns.

Infrastructure Use Case Library

The framework includes a scored library of 30+ data infrastructure use cases spanning governance implementation, data quality remediation, integration architecture, streaming pipeline deployment, MLOps standardization, vector search implementation, and agent infrastructure buildout. Each use case maps to specific EBITDA impact ranges, prerequisite capabilities, implementation timelines, and compound effects on downstream AI initiatives.

Conclusion: Infrastructure Is the Investment

The data infrastructure market in 2026 presents a paradox that is also an opportunity. Enterprise AI investment has reached $644 billion, yet only 14% of organizations can confidently measure returns. Hyperscaler capex is approaching $700 billion annually, yet the primary constraint is not compute but electricity. Databricks has achieved a $134 billion valuation on $4.8 billion in revenue, yet 60% of enterprise AI projects fail because the data feeding them is not ready.

The resolution of this paradox is the same insight that Sharp Business Systems learned the hard way: the AI model is never the bottleneck. The data infrastructure is. Organizations that invest in governance, quality, integration, and real-time processing before deploying AI achieve three times faster deployment with 60% higher success rates, 40% higher analytics ROI, and measurable revenue and cost improvements that compound with each subsequent initiative.

For PE operating partners, the implication is that data infrastructure is not a technology line item to be managed. It is the investment that determines whether every other AI investment in the portfolio produces returns. Blue Orange Digital's AI Data Optimization Framework provides the diagnostic and implementation methodology to make that investment with precision: assessing readiness, identifying gaps, sequencing initiatives by compounding impact, and measuring results against quantified baselines. In a market where 86% of enterprises cannot measure AI ROI, the ability to consistently produce measurable returns from data infrastructure investment is the most defensible competitive advantage a PE fund can build.

Ready to Accelerate AI Value Creation?

Blue Orange Digital partners with PE operating teams and portfolio companies to design, build, and scale AI data systems that deliver measurable EBITDA impact.

blueorange.digital | **hello@blueorange.digital**

About Blue Orange Digital

Blue Orange Digital is a data engineering and AI consultancy specializing in building production-grade AI systems for private equity-backed companies. We combine deep vertical expertise with proven technical frameworks to accelerate value creation across the portfolio.

New York | Washington DC | **blueorange.digital**