Futurum

Google Cloud

Escaping Data Gravity and Infrastructure Debt: Why the AI Era Demands an Agentic Data Cloud

A pragmatic analysis of how enterprises are reclaiming human capital and slashing operational overhead by migrating from decoupled data lakes to natively integrated, in situ reasoning engines.

AUGUST 2026

The Gravity Problem: Reassessing Data Infrastructure for Generative Workloads

Boardrooms are handing technology leaders a clear edict: integrate generative artificial intelligence and autonomous agents into daily operations immediately. And companies are listening. According to the Futurum Research 1H 2026 Data Intelligence, Analytics, and Infrastructure Decision Maker Survey Report, 74.2% of large enterprise technology leaders state they have a well-formed plan for how Agentic AI can achieve measurable operational efficiencies. Yet, as organizations move from isolated conversational pilots to production-grade autonomous systems, physical infrastructure is becoming a massive obstacle.

Futurum Research Note

This report combines:

- Futurum Research quantitative survey findings
- Independent analysis of the market
- Primary interviews conducted by Futurum Research with enterprise technology leaders who have recently modernized or migrated their analytics environments

Customer perspectives presented throughout the report are anonymized interview excerpts used with participant permission.

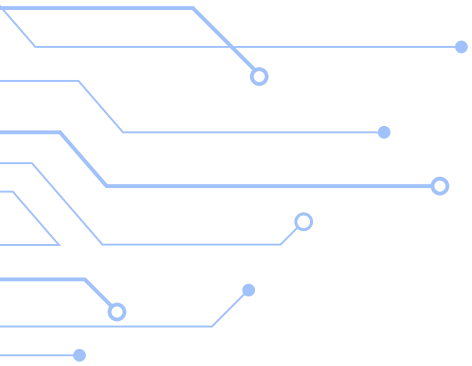
Modern AI agents demand rich, dynamic, and semantically grounded context to deliver accurate reasoning. The vast majority of enterprise data, however, remains split across a deep operational and analytical divide. Crucial transactional and operational data is trapped in disparate databases, ERPs (e.g., SAP or Oracle), and SaaS platforms (e.g., Salesforce), separated from unstructured data and analytical lakehouse environments. Traditionally, bridging this divide required continuously replicating data from operational systems and applications into analytical systems. By the time this transactional and live application data is extracted, processed, and landed, it is already stale. For real-time agentic AI, this data latency breaks the reasoning loop, depriving agents of the immediate operational context they need to act.

When a business application asks a model to make a decision, that model needs recent, specific context to return a useful result. Packaging large volumes of information, sending it across a network, and loading it into an outside system takes time and adds complexity.

Autonomous AI agents cannot make accurate, operational decisions based on 12-hour-old batch-processed data. The agentic era demands continuous, real-time context. An effective move to agentic workloads will require organizations to completely blur the lines between analytical and operational workloads. Data processing engines must support continuous queries and real-time streaming natively, meaning when an AI agent evaluates a business state or triggers a workflow, it must be able to reason over the enterprise's reality as it exists at that very second.

This physical separation of the Analytical and Operational data estates also fractures enterprise security. The strict privacy rules that control who can see specific records in the main database rarely transfer when teams create a copy for an external tool, leaving sensitive records vulnerable. This challenge will rapidly multiply as autonomous agents are deployed and work across enterprise data ecosystems. Beyond the security risks, cloud providers charge fees to move large datasets across their networks, which quickly drives up project costs. Ultimately, this fragmented setup forces engineering teams to act as digital plumbers, spending their weeks fixing broken connections and monitoring transfers rather than building the intelligent systems the business actually needs.

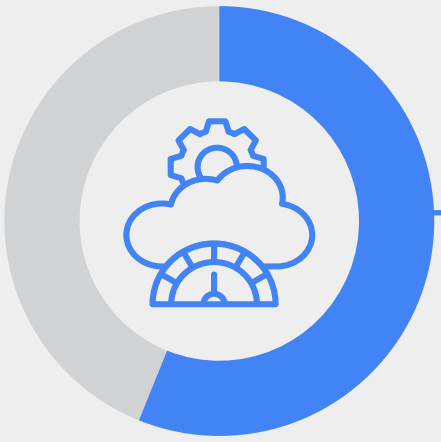
For many companies, data catalogs promise to provide grounding context. However, traditional catalogs were built as passive repositories for human specialists rather than active unified context systems for AI. Legacy catalogs excel at indexing database schemas, but they fail to capture the rich semantic relationships and business logic often embedded in the company's vast unstructured data assets, and agents need that context to make decisions. Without a living context engine that translates raw technical metadata into real-world business meaning, agents are left to interpret data in isolation, ultimately leading to hallucinated reasoning, sluggish response times, and broken workflows, thereby eroding human trust.



Across Futurum Research's primary interviews with enterprise technology leaders who recently migrated from Databricks to BigQuery, respondents reported an average 21% reduction in TCO and a 32% decrease in operational overhead within the first few quarters of deployment.

Driven by these issues, viewed against the backdrop of massive inference costs and the undeniable physics of data gravity, IT leaders are completely rethinking their baseline architectures (see Figure 1). According to the Futurum Research 2026 Key Issues & Predictions report, 71% of CIOs are actively reevaluating their cloud workload placements, moving aggressively away from fragmented infrastructure toward highly integrated, serverless environments.

Figure 1: The Urgent Enterprise Pivot in Cloud Workload Placement



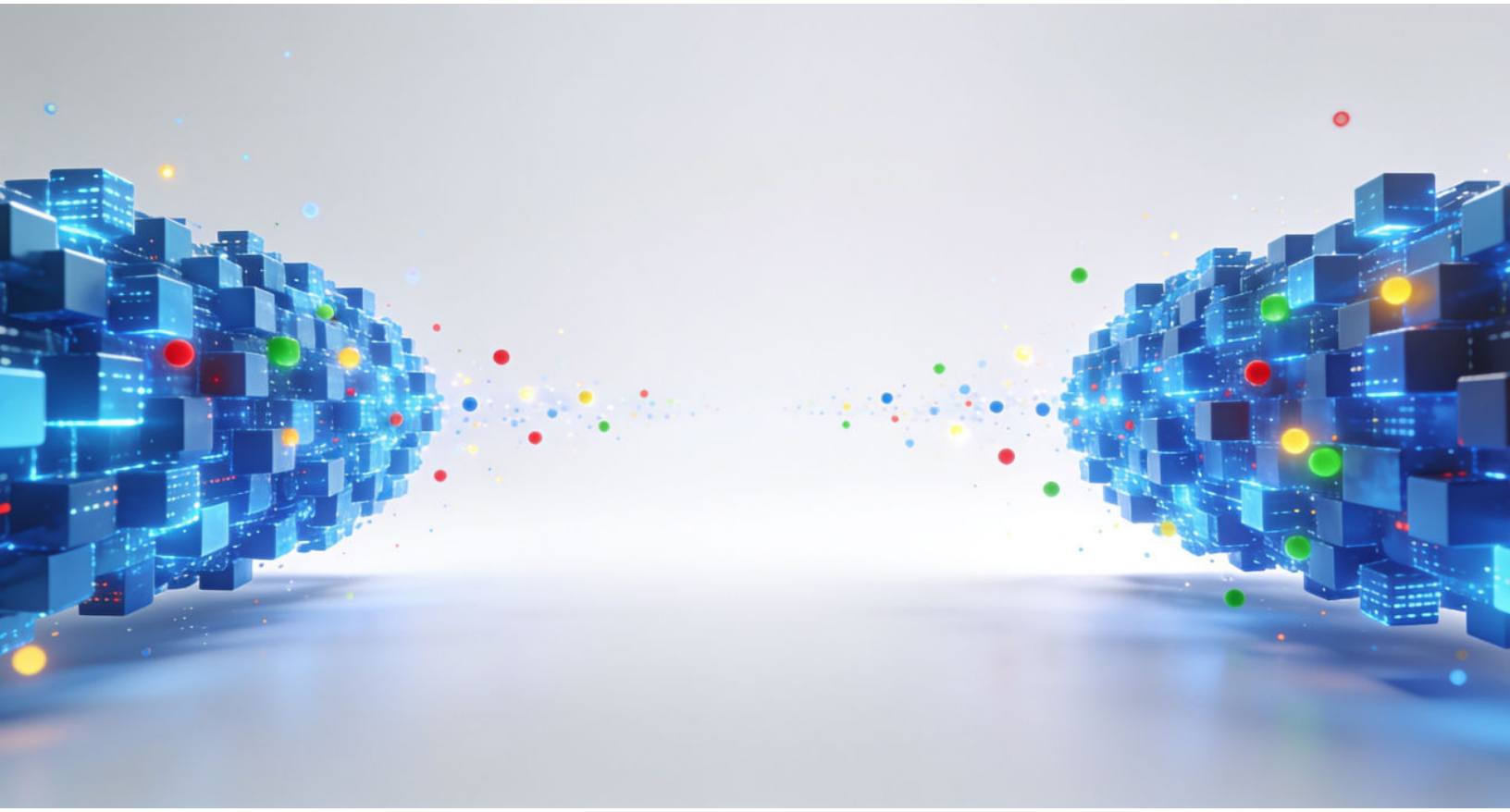
56.1%

Actively reevaluating IT workloads across public clouds and on-premises

Compelled by the demanding economics of artificial intelligence and the reality of data gravity, the majority of large enterprises are rethinking their workload placements to find integrated, cost-effective environments that eliminate external data transfers.

Base: n=244 large enterprise technology leaders ($1B+ in revenue)
Futurum Research, 1H 2026 Data Intelligence, Analytics, and Infrastructure Decision Maker Survey Report.

Q: "What are your primary data storage and workload placement strategies for the next 12-18 months?"



The Hidden Financial and Human Toll of a Decoupled Stack – From Infrastructure to Analytics and AI

To bridge the gap between static storage (where data sits idle in warehouses and lakes) and business activation (where data science, machine learning, AI, and BI occur), organizations spent the last half-decade heavily investing in external, third-party computing platforms. While these vendors pitch their platforms as 'integrated', the reality is a deeply decoupled architecture: foundational storage remains structurally separated from the complex compute layers required for real-time activation, imposing an unsustainable integration tax on the modern business.

Even when deploying these platforms under the guise of an 'all-in-one' lakehouse, organizations face a hidden infrastructure tax. Because third-party compute engines rent their underlying infrastructure from hyperscalers, they often pass premium compute unit costs and cross-network egress fees directly onto the customer. Moving data from secure analytical storage to external, decoupled compute clusters for ML modeling or agent grounding inherently inflates the total cost of ownership (TCO) and introduces unacceptable latency into real-time pipelines.

Customer View

"We had to stand up compute layers throughout the duration, Monday to Monday, 9 am to 9 pm, at the same level. We were bleeding on compute layers if we were not using it."

– Director of Software Engineering, Multinational Financial Services Firm

These decoupled architectures also drain human capital. Highly skilled data engineers find themselves forced into the role of cluster administrators. They spend their days guessing capacity requirements, tuning infrastructure for peak workloads, and monitoring idle compute cycles over the weekends. Furthermore, while platforms such as Databricks and Snowflake heavily market their catalogs (such as Databricks Unity Catalog or Snowflake Horizon) as 'open' frameworks, there is additional context to be considered. While their basic storage formats might be open, the advanced governance, lineage, fine-grained security, and cataloging features enterprises actually require to run safely are kept strictly proprietary. This traps organizations in a commercial management and business-critical governance layer that only grows in complexity over time and cannot be ported to other platforms, locking down their data gravity.

Compounding this operational exhaustion, enterprise skills shortages have increased 5.6 percentage points to reach 10.4%, officially surpassing available budget as the most critical constraint to AI adoption, according to Futurum Research. Organizations simply lack the engineering bandwidth to waste on manual infrastructure maintenance and learning disparate systems and technologies.

Furthermore, routing data outside the secure warehouse boundary dismantles existing role-based access controls. Once sensitive enterprise intellectual property leaves the native database to enter a third-party processing layer or external vector database, organizations expose themselves to significant privacy risks and compliance violations (see Figure 2).

Figure 2: The Security and Governance Roadblock in Production Architecture

Base: n=244 large enterprise technology leaders ($1B+ in revenue)
Futurum Research, 1H 2026 Data Intelligence, Analytics, and Infrastructure Decision Maker Survey Report.

Q: "Which of the following data management and governance hurdles are most critical to resolve before launching production AI?"

Architecting for Autonomy: Google Cloud's Triad of Value

Resolving the friction between data gravity and the demands of AI requires a converged architectural approach. Instead of moving massive data blocks closer to external models, enterprises must run secure, governed processing directly where the information natively resides. Google Cloud's BigQuery, part of the company's Agentic Data Cloud, addresses this directly by establishing a completely serverless, vertically integrated data and AI foundation designed specifically for the era of agentic AI.

Data infrastructure providers are moving away from decoupled architectures by embedding native AI capabilities directly into the underlying storage and database layers. To execute this transition successfully, the BigQuery architecture rests on three competitive pillars.

SQL Workloads: The Serverless Bread and Butter

BigQuery eliminates the concept of cluster tuning and capacity planning entirely. By operating natively as a fully managed, serverless platform, it scales compute resources instantaneously on a per-query basis. This removes the operational overhead of managing third-party clusters and stops the financial bleed of paying for idle infrastructure. Data teams pay strictly for the exact queries they execute, bringing total predictability to the cloud budget while freeing engineers to focus on business logic. For massive SQL workloads, this means unparalleled price-performance without the babysitting.

Moving to a serverless model solves the engineering headache of cluster management, but it frequently introduces a new tension with finance teams. When organizations pay per query, a poorly written line of SQL by a junior analyst can result in an unexpectedly high daily cloud bill. The practical middle ground that enterprises are adopting involves using BigQuery's capacity management features. By reserving a baseline of compute slots for predictable daily workloads and relying on on-demand scaling only for unexpected spikes, teams can bridge the gap between engineering speed and financial predictability. This changes the conversation from reacting to surprise costs to actively managing compute as a measurable utility. With BigQuery fluid scaling, you can further optimize price-performance and costs when executing highly variable workloads with a premier autoscaling model that does not require a cost-and-performance trade-off. Fluid scaling in BigQuery enables true per-second billing, offering up to 34% cost savings.

Apache Spark Workloads: Open Portability and Peak Performance

Enterprises are understandably hesitant to abandon heavy historical investments in Spark, often viewing legacy lakehouse environments as too complex to untangle. Google Cloud resolves this by offering complete architectural flexibility: organizations can run SQL workloads in BigQuery or leverage Managed Service for Apache Spark for dedicated Spark engineering. Fully compatible with open-source Spark, Managed Spark is powered by Lightning Engine, Google Cloud's native, high-performance execution engine that delivers up to 4.9x the performance of open-source Spark, and up to 2x the price-performance over the leading high-speed Spark alternative. Managed Service for Apache Spark is available in both serverless and cluster-based deployment options. Together, these interoperable, massively scalable engines allow data teams to retain their preferred frameworks without operational complexity.

Google's deep integration with open table formats such as Apache Iceberg neutralizes the threat of vendor lock-in by interoperating with BigQuery as well as open source engines such as Apache Spark without moving data or worrying about read/write interoperability. Storage remains decoupled from compute structurally, allowing data teams to retain their preferred engineering frameworks and open cataloging tools without suffering the operational bloat of legacy cluster management.

Customer View

"We didn't just shift the platform. We shifted from platform management to value delivery."

– Enterprise Architect, Global Technology Consulting Firm

The adoption of open table formats such as Apache Iceberg represents a structural change in how companies negotiate with cloud providers. Historically, once data was loaded into a proprietary data warehouse, data lake, or lakehouse, moving it to another provider was financially unfeasible. By decoupling storage from compute and keeping the underlying data in an open format, enterprises maintain the flexibility to point different processing engines at the same data without copying it. This changes the dynamics for platforms such as Google Cloud's borderless Lakehouse, which connects open formats such as Apache Iceberg, Delta Lake, and Apache Hudi to engines such as BigQuery and Managed Service for Apache Spark, emphasizing differentiation through query execution speed and price, rather than relying on storage lock-in to retain customers.

Native AI Integration: The Differentiator

The most critical advantage of the BigQuery platform is its natively embedded intelligence, that is, AI built in, rather than bolted on. Google owns both the underlying first-party foundation models (e.g., Gemini) and the processing platform Gemini Enterprise Agent Platform. This allows organizations to execute complex models directly against their data via standard BigQuery AI and AI functions in BigQuery to help with AI tokens estimations and controls. Running models in situ satisfies the enterprise security mandate natively, preserving strict row- and column-level access controls without ever piping data to an external service.

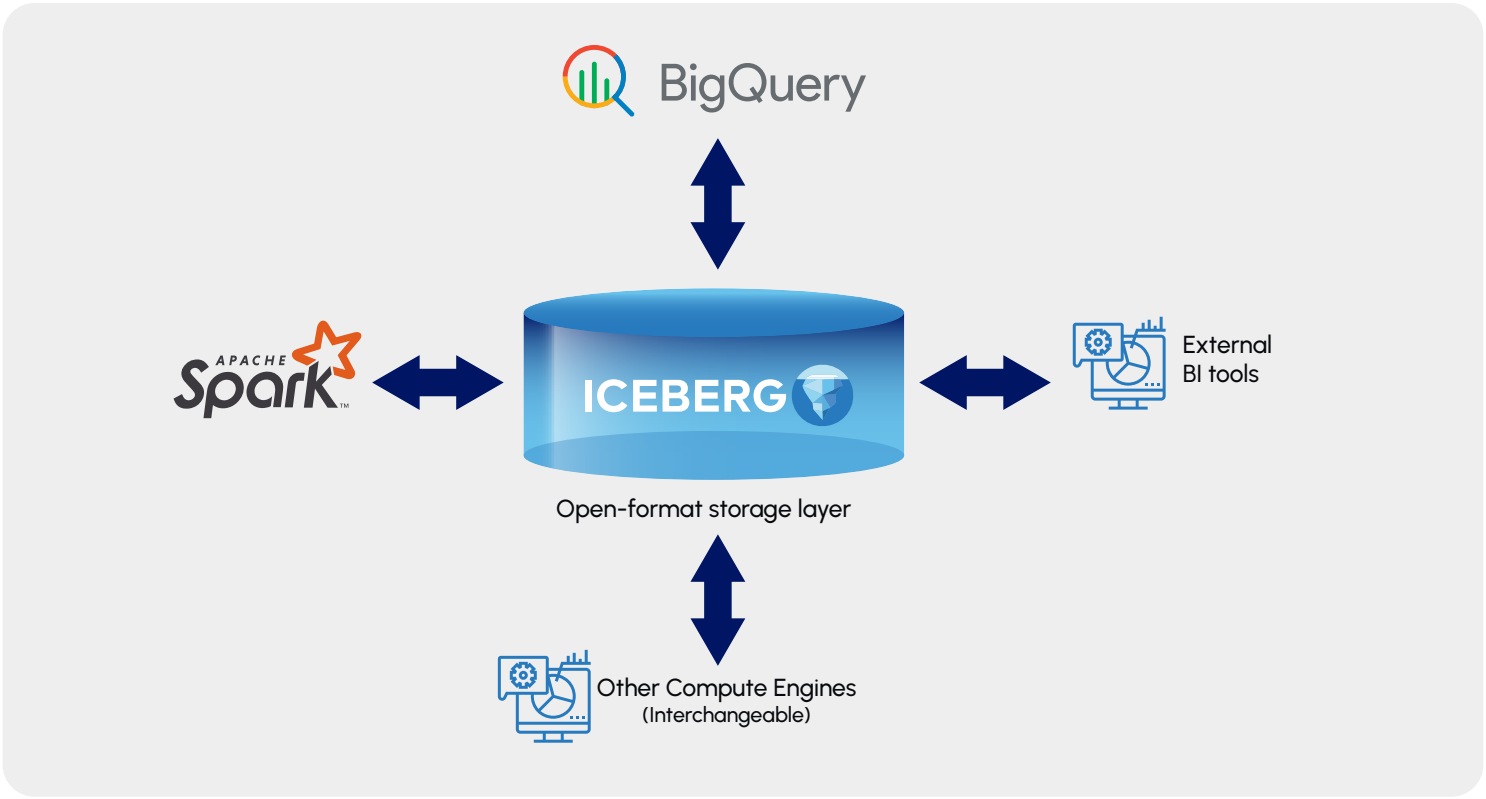
To illustrate, BigQuery, which is an integral part of the Agentic Data Cloud, natively unifies structured tables with unstructured enterprise data (e.g., PDFs, audio logs, and support transcripts). Enterprises can use native AI functions in BigQuery to process and analyze unstructured data using SQL, simplifying tasks such as document parsing, sentiment extraction, semantic chunking, vector search, and more. By leveraging deep integration with Gemini and Gemini Enterprise Agent Platform, autonomous agents can query and reason across multi-modal data in a single sweep. Enterprises no longer need to maintain separate, expensive database silos just to process text and images alongside their transactional rows. Furthermore, enterprises can run zero-shot inferencing using cutting-edge pre-trained models from Google Research and DeepMind, such as TimesFM for time series forecasting, TabularFM for tabular data classification and regression, and WeatherNext for weather forecasting.

Beyond immediate security concerns, physical data movement fractures the audit trails required for regulatory compliance. As regions implement stricter guidelines on how artificial intelligence is trained and deployed, organizations are required to prove exactly which datasets informed a specific model decision. When data is extracted from a central warehouse and piped into an isolated AI environment, that chain of custody breaks. Running model inference natively within the database allows compliance teams to maintain a continuous, verifiable record of data lineage from the initial input down to the model's final output.

From Friction to Fluidity: Migration Mechanics and Real-World Velocity

Transitioning from Spark-heavy environments such as Databricks to a fully managed Google Cloud infrastructure yields immediate financial and operational dividends. Because Google owns the entire stack, from the hardware on up to the AI model, the hidden penalties of data movement and infrastructure management vanish. Organizations migrating their workloads consistently report substantial drops in their total platform costs, driven almost entirely by the elimination of third-party egress fees and idle compute charges (see Figure 3).

Figure 3: The End of Storage Lock-In



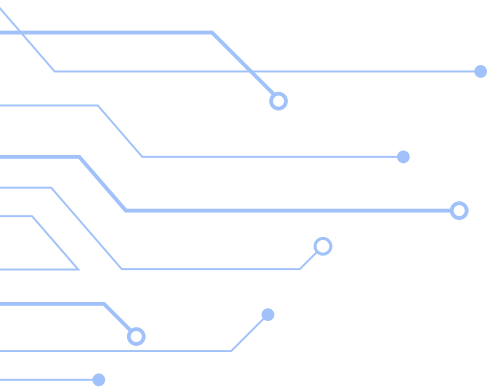
Customer View

"The main cause that we had was egress-ingress out of various cloud platforms...just pulling data and then pushing it was costing a lot of money."

– Director of Software Engineering, Multinational Financial Services Firm

Equally important is the reality of the migration process itself. Historically, rewriting decades of custom pipeline code stood as the primary barrier to modernization, keeping companies locked into legacy ecosystems for fear of downtime. Today, enterprises are utilizing highly capable large language models (including Gemini and Claude) to autonomously translate historical codebases into optimized native SQL and serverless Spark deployments. This sort of AI-assisted velocity can reduce migration timelines from years to mere months.

Customer View

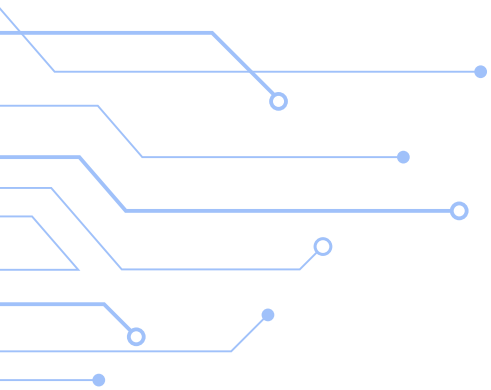


"We heavily [use] AI to do this one. It's not someone who is doing [it] from scratch...we were able to gain around 30% to 50% productivity gain by leveraging AI on this translation."

– Enterprise Architect, Global Technology Consulting Firm

Once stabilized within Google's Agentic Data Cloud, data teams find themselves operating in a fluid ecosystem rather than a rigid pipeline. BigQuery acts as the central reasoning engine, natively interoperating with the open lakehouse, serverless Spark, real-time streaming data, and operational databases (such as Cloud SQL and AlloyDB). This eliminates the friction of moving data between transactional and analytical silos (see Table 1). Data scientists, business analysts, and data engineers can collaborate within the exact same environment. Google Cloud Lakehouse further expands this borderless capability, allowing organizations to run cross-cloud queries directly against AWS or Azure footprints. This interoperability is what actually enables autonomous AI agents to reason across the entire enterprise estate (everything from historical lakehouse files to up-to-the-second streaming events) safely within a governed boundary.

Customer View



"It's accelerated [artificial intelligence] adoption by bringing [machine learning] directly to the data."

– Enterprise Architect, Global Technology Consulting Firm

Table 1: Decoupling from Legacy Friction – The Migration Impact Matrix

Enterprise Pain Points	BigQuery's Architectural Solution
The Data Movement Tax: Third-party analytics platforms require constant egress and ingress across physical cloud networks, aggressively inflating monthly cloud consumption bills.	**Vertically Integrated Economics:** Because Google owns the underlying network fabric and hardware and offers capabilities such as a high-speed interconnect and cross-cloud caching, the hidden costs of external data movement are structurally eliminated.
Operational Exhaustion: Highly skilled data engineers burn endless cycles guessing capacity needs, managing clusters, and paying for idle weekend compute.	**Fully Serverless Execution:** True auto-scaling serverless infrastructure removes capacity planning entirely, allowing organizations to pay strictly for executed queries.
Bolted-On Security Risks: Extracting secure enterprise data into external vector databases breaks role-based governance and risks intellectual property exposure.	**Native In Situ Vector Search:** Google's Agentic Data Cloud brings vector search and indexing natively into BigQuery and its operational databases (e.g., Cloud SQL and AlloyDB). This deep integration allows organizations to build retrieval-augmented generation (RAG) pipelines and run inference directly against their data securely inside the warehouse boundary. It completely eliminates the need to pipe secure IP information into an external vector database stack.
Legacy Code Lock-In: Migrating years of custom Spark pipelines and complex catalogs is viewed as too risky and resource-intensive for resource-constrained data teams to undertake.	**AI-Assisted Migration Velocity:** The utilization of highly capable language models enables rapid, automated code translation from legacy formats to optimized SQL.

Strategic Imperatives for the AI-Ready Data Enterprise

Modernizing the enterprise data stack is no longer an exercise in chasing marginal query speed improvements. It is a fundamental requirement for clearing the operational friction that prevents human talent from securely deploying AI models and agents from operating on your enterprise data. To execute a successful migration and establish an AI-ready foundation, data leaders should adopt the following tactical imperatives.

Audit the Invisible Taxes of Data Movement: Before initiating a platform change, calculate the exact financial damage caused by current egress, ingress, and idle compute charges. Establishing this baseline is essential for proving the return on investment of a vertically integrated, serverless solution.

Automate the Migration Workflow: Refuse to manually rewrite thousands of lines of historical pipeline code. Leverage modern generative AI tools and specialized integration partners to rapidly translate legacy workflows into highly optimized, native SQL.

Mandate In Situ Processing for Corporate Security: Treat any architectural proposal that forces you to pipe sensitive data outside the core governance boundary as a fundamental security flaw. Consolidate predictive and generative models directly adjacent to the data warehouse to maintain strict role-based access controls.

Reallocate Engineering Talent to Value Creation: Use the transition to a serverless architecture as a forcing function for a broader cultural shift. Move the brightest data engineers away from cluster babysitting and infrastructure maintenance, tasking them instead with building the autonomous, intelligent workflows that drive tangible business revenue.

The objective of data modernization has fundamentally evolved. Organizations must stop building complex bridges between their data and their intelligence. By standardizing on a highly integrated, serverless foundation such as BigQuery and Google's Agentic Data Cloud, enterprises can finally stop managing platforms and start generating substantive business value.

Learn how Google Cloud can help. Request a data strategy workshop today.

Important Information About This Report

Authors

Brad Shimmin
Vice President & Practice Lead, Data Intelligence, Analytics, & Infrastructure | The Futurum Group

Publisher

Futurum Research

Inquiries

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

Citations

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

Licensing

Disclosures

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Google Cloud

About Google Cloud

Google Cloud is Google's suite of cloud computing services, offering infrastructure, storage, machine learning, and data analytics tools that businesses use to build, scale, and run applications. A standout offering within the platform is BigQuery, a fully managed, serverless data warehouse that lets organizations run fast SQL queries across massive datasets without managing any underlying infrastructure. BigQuery integrates seamlessly with other Google Cloud services and supports real-time analytics, making it a popular choice for companies looking to derive insights from large volumes of data quickly. Beyond BigQuery, Google Cloud also provides compute (via Compute Engine and Kubernetes Engine), AI and machine learning tools (like Vertex AI), and robust security and networking capabilities. Overall, Google Cloud competes with AWS and Microsoft Azure as one of the major hyperscale cloud providers, with BigQuery often cited as one of its most differentiated and widely adopted products.

Futurum

About The Futurum Group

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.