OCTOBER 2025

Illuminating Dark Data With Smart Storage from Google Cloud

Simon Robinson, Principal Analyst

Abstract: Organizations today struggle to extract meaningful insights from the vast repositories of unstructured "dark data" that they routinely store and retain, often for years—a huge, missed opportunity. Traditional approaches to data management suffer from relying on human-driven processes that cannot scale to handle petabyte-level volumes, creating bottlenecks in preprocessing pipelines and requiring subject matter experts to define semantic meaning before queries can be run. By contrast, Google Cloud's smart storage platform addresses this challenge by embedding AI capabilities directly into the storage layer, transforming it from a passive repository into an intelligent analytical system that can automatically process and derive insights from rich media content. By leveraging a vertically integrated technology stack that includes Google Cloud Storage, Gemini LLM, BigQuery, and Vertex AI, organizations can now unlock new value from their entire data estate with unprecedented levels of automated analysis.

The Dark Data Challenge

A major consequence of today's digital age is the tremendous volumes of digital data that we create. Whether in our personal or business lives, the volume of data we generate—especially unstructured data such as camera-generated images and video—continues to explode. Indeed, unstructured data accounts for the majority of enterprise data at most organizations; according to research from Enterprise Strategy Group, unstructured data accounts for at least 61% of total data at almost three-quarters of organizations, and for a significant minority is substantially higher.[1] Much of this data is retained, especially in cloud-based archives—most of it indefinitely—and much of this data is rarely or never accessed.

Failing to leverage this enormous repository of dark data represents a huge, missed opportunity to reveal meaningful insight and potentially create new value. Despite the emergence of "big data" over the last decade or so, we are still only scratching the surface of the mountain of data we are sitting on; many companies never use even a tiny fraction of what they store. Indeed, many IT leaders don't fully understand what they are storing: two-thirds of organizations said they regularly encounter issues with visibility into their data.[2]

There are many reasons for this, including poor quality and inconsistent metadata, multiple inaccessible or incompatible data formats, and siloed systems that make data difficult or even impossible to query. However, at the core of these issues is a common challenge: a lack of understanding at a system level of the nature of the data being stored.

Instead, understanding data is still largely a human-driven process, which introduces two problematic bottlenecks. First, preprocessing pipelines to cleanse and separate data are still largely manually built. And second, subject matter experts are required to define the semantic meaning and context of data before a query can be run. With

[1] Source: Enterprise Strategy Group Research Report, *Reinventing Data Loss Prevention: Adapting Data Security to the Generative AI Era*, May 2025.
[2] Source: Enterprise Strategy Group Research Report, *The Critical Role of Storage in Building an Enterprise AI Infrastructure*, September 2025. All Enterprise Strategy Group research references in this Showcase have been taken from this report unless otherwise noted.

many organizations routinely storing petabytes of unstructured data, it is simply impossible to scale human operations to support such data volumes beyond the most obvious, high priority data sets.

As a result, many organizations find they are incurring the cost and risk of storing vast amounts of data without the ability to unlock the insights that might exist within. In addition, they are missing out on a huge opportunity, to take advantage of their dark data in a way that could drive meaningful benefits for the broader organization. For example:

- Avoiding a security compromise by recognizing potentially damaging patterns and behaviors in surveillance camera video footage.
- Avoiding inadvertently displaying fake or AI-generated products on an e-commerce website.
- Reducing or eliminating high model bias in training models due to minimizing the need for human-in-the-loop annotation or a random selection of data to compose data sets.
- Advancing rollout of advanced autonomous driving capabilities by reducing model training timeframes.
- Improving the social media platform experience by reducing the volume or eliminating offensive, objectionable, or inappropriate content.

Enter AI: Unlocking Value From Dark Data

The rapid emergence of AI technologies—and generative AI (GenAI) in particular—presents an opportunity to unlock the value hidden within dark data. LLMs excel at interpreting and generating natural language, enabling them to derive context, meaning, sentiment, entities, and complex patterns directly from free-form text and media. Such capabilities promise to address the stubborn bottlenecks that have limited an organization's ability to understand its dark data at scale, in the process eliminating brittle preprocessing pipelines and manual schema definitions. Instead, the model can inherently understand raw data on its own.

There's a catch, however. Foundational models are trained on public data, not organizations' own data. This presents IT leaders with multiple additional challenges:

- **Data discovery and preparation.** This initial phase is crucial to AI success, and avoiding the 'garbage in, garbage out' problem. However, most enterprise data resides in passive storage environments that have little understanding of the nature of the data they are storing. This means data professionals must spend time and effort building additional systems to extract more detailed understanding of the data contained within storage environments.
- **AI data pipeline/workflow integration.** The multiple tools and capabilities that enable GenAI workflows across the entire pipeline—data repository, foundational models, structured knowledge platform, knowledge graphs—have only until recently been available from separate vendors. This places the emphasis on the customer (or a partner) to do the pipeline integration. In an environment of extreme AI skill shortages, this is a significant challenge that drives up costs, elongates project timelines and ultimately reduces the ROI of any AI initiative. Seventy-one percent of organizations reported that effectively integrating storage with data pipelines and AI workflows is a major challenge.
- **Data security and privacy.** Organizations need to be confident their data will be used in a compliant manner that protects their sensitive data across the entire AI lifecycle. These considerations are particularly relevant in the context of dark data which, as well as being large and fast growing, is also unwieldy and, to a large extent, unknown.

It's clear that many organizations stand to benefit substantially from deeper, automated insight derived from their large volumes of unstructured data—both historical and ongoing. However, to tackle this effectively, IT leaders will need to develop a strategic response that leverages emerging technologies such as GenAI but in a comprehensive manner that considers the end-to-end AI data pipeline, as well as enterprise considerations. Such a solution would be able to:

- Securely leverage an organization's entire data estate, including all data types that increasingly include rich unstructured data and media.
- Effectively and securely integrate all key elements of the GenAI lifecycle into a cohesive process that minimizes complexity, reduces risk, and increases business value and ROI by unlocking insight into all relevant unstructured data sources.
- Deliver at scale, as data volumes are growing to unprecedented degrees, to the extent that applying new intelligence to data in an automated manner is the only way to cost-effectively scale.

Google Cloud's Differentiation: A Vertically Integrated Cloud Stack

Google Cloud's strategy in AI is to create a vertically integrated technology stack that can enable a customer's entire AI journey end to end. This unique approach has helped establish Google Cloud as a leader in the AI space, enabling its customers to seamlessly connect their private data directly to foundational intelligence.

Google Cloud can do this because it owns best-in-class capabilities across the AI pipeline, offering substantially differentiated features at every stage, combined with integration across all elements, which vastly simplifies the customer experience, reduces risk, and shortens time to value.

The key elements of the Google Cloud stack include the following:

- **Data repository:** Google Cloud Storage is Google's core storage offering, providing highly scalable, reliable, and secure unstructured data store.
- **Foundational models:** Gemini is Google's LLM, providing high-quality AI capabilities across multimodal data sets, including audio, images, software code, text, and video.
- **Unstructured knowledge platform/data warehouse:** Google BigQuery is a fully managed, AI-ready data platform that helps customers manage and analyze their data with built-in features, including machine learning (ML), search, geospatial analysis, and business intelligence.
- **A unified AI platform:** Vertex AI is a unified ML platform from Google Cloud that provides tools and services to build, train, deploy, and customize ML models and AI applications, including GenAI. Vertex AI streamlines data engineering, data science, and ML engineering workflows, offering options from no-core AutoML to full custom model training, and includes services such as Model Garden for testing GenAI models, Vertex AI Studio for rapid prototyping, and Vertex AI Agent Builder for creating conversational AI agents.

One final aspect worth noting is the vast amount of experience Google brings through building, scaling, and enabling AI in several of the world's largest unstructured digital data repositories, including YouTube and Google Photos. This experience has provided it with essential insight into the specific data engineering and infrastructure considerations when architecting advanced AI capabilities to exabytes of rich video and image-based content. All of these learnings have been incorporated into Google Cloud's strategy to help other organizations bring value to their dark data.

Google Cloud's Vision for a Smart Storage Platform

Google Cloud is now further extending this vertically integrated approach by embedding additional, innovative capabilities into the storage layer itself: the smart storage platform. The platform promises to transform storage from a passive layer that stores blobs into an intelligent, analytical layer that can unlock new value, both from troves of dark data as well as from new data as it lands in the system. However, for Google Cloud, this isn't about doing the same things in a better way but rather unleashing brand-new capabilities previously unavailable to organizations.

To achieve this, Google Cloud decided to address the most difficult aspects of data automation with smart storage, specifically by applying these new capabilities to rich media content types, including images and video. This is truly groundbreaking work. For the most part, LLMs to date have been trained using text-based data; training models on

video and image data is much more difficult, and Google is the first to make this a reality *at scale* for enterprise use cases.

The ability to apply these technologies to gigantic data stores that can reach petabyte or even exabyte levels of scale is critical. The limitations of human-centric processes are really exposed at these volumes because humans can only manually classify or touch a tiny fraction of data; by contrast, Google Cloud's smart storage does the heavy lifting across all data, enabling staff to focus on other tasks or apply their expertise in a more focused, value-added manner.

As a result, smart storage will empower data engineers or scientists and business users to gain new insights, drive new business value, mitigate risks, and even unleash new revenue streams from all of their unstructured data with unprecedented levels of analysis.

The smart storage platform has been designed to automatically activate an AI-driven pipeline, generating the right vector embeddings for semantic search, extracting key relationships into a graph database, and making all unstructured data automatically and immediately available to higher-level AI tools and agentic systems. By leveraging the full power of Google Cloud's ecosystem, raw data is transformed into actionable, governed insights with minimal to zero manual intervention from the customer. Crucially, the entire process is governed by customer permissions, ensuring that the right data is utilized at every stage.

Unleashing Value From Data Across Multiple Use Cases With Google Cloud Smart Storage

A key aspect of Google Cloud's smart storage platform is the ability to apply it to a broad set of use cases as part of a broader Google Cloud AI workflow. Since the data has already been vectorized or processed, with the knowledge graph built and semantic meaning already understood, users can ask complex questions in natural language. For example:

- If a security firm is monitoring large volumes of video footage daily, rather than having to go in and review data after the fact to identify relevant information (e.g., scenes with people, weapons, etc.), users can simply ask the system to review all footage from location X in the last 48 hours and identity all scenes with people.
- E-commerce marketplaces might struggle to police their product catalog for policy violations. Their auditing processes can better scale with proactive alerting using AI/ML against their unstructured data to highlight policy or compliance risks for human-in-the-loop investigation.
- Autonomous driving cannot afford failure, so the supported AI/ML models depend upon addressing gaps to eliminate risk by finding needle-in-a-haystack data. Users can ask the system to find these rare but critical pieces of data and greatly accelerate their model evolution.
- Social media platforms deal with challenging security and privacy concerns, and problematic content will inevitably slip through the first line of defense. Smart storage can support a more thorough second line of defense by applying more powerful AI/ML models at scale and automatically quarantining risky content that needs human review.

> "With the latest AI innovations, we are transforming Google Cloud Storage into a platform that not only stores data but also gives meaning to every byte to empower enterprises to use 100% of their data for critical business decisions. With auto annotate and Object Context, we are delivering the initial two pillars of our vision. This is the first step toward making Google Cloud Storage a truly smart storage platform."
>
> – Asad Khan, Sr. Director Product Management at Google Cloud

Google Cloud recently announced two new important capabilities as part of its smart storage strategy. Working together, they are designed to deliver automatic, rich insights and unlock instant discovery across a customer's massive data estate.

"Our customers trust us to help make their homes and lives safer, smarter, and more convenient, and AI is at the heart of our product and customer experience innovations. Cloud Storage auto annotate's rich metadata delivered in BigQuery helps us scale our data discovery and curation efforts, speeding up our AI development process from 6 months to as little as 1 month by finding the needle-in-a-haystack data essential to improve our models."

– Brandon Bunker, VP of Product, AI, Vivint

- **Auto annotate:** Designed to illuminate dark data with an initial focus on images, auto annotate automatically generates rich metadata from images using Google's Advanced AI models. Auto annotate was designed to be simple and can be switched on for any or all of a customer's Google Cloud Storage buckets. Users can simply choose the model to be applied, and the entire image library will be annotated, with new images automatically annotated as they are uploaded. When auto annotate is connected to a data warehouse, all annotations are made available to BigQuery, which enables customers to use BigQuery Vector Search across the entire storage estate. The initial version of auto annotate will offer three models: object detection with bounding boxes, image labeling, and objectionable content detection with score.

- **Object Context:** Now available in preview, this is Google Cloud's foundational feature for building a truly smart storage platform. With it, customers can attach custom key-value pair metadata with creation and modification timestamps directly to objects, making data searchable, actionable, and rich with lineage information. Object annotations are also available as contexts, which are integrated with Google Cloud Storage APIs to provide a flexible and native way to enrich data and interact with it. Combined with smart storage features like auto annotation, which generates contexts at scale, Object Context converts data into information to enable sophisticated data management workflows directly within the storage layer.

"Object Context gives us a way to take the outputs of BigID's industry-leading data classification solutions and apply labels to Cloud Storage objects. Object Context will allow BigID labels to shed light onto data in Cloud Storage: identifying objects which contain sensitive information and helping them understand and manage their risk across AI, security, and privacy."

– Marc Hebrard, Principal Technical Architect, BigID

Conclusion

The AI revolution continues to open all manner of opportunities for forward-thinking organizations to drive new value, efficiencies, and even revenue streams from their data. As the volume of unstructured data continues to explode at organizations globally, a major new opportunity is emerging to leverage this data at scale in a more fundamental and meaningful manner.

However, bringing light to vast reserves of dark data at compelling levels of scale requires a comprehensive approach that blends a range of leading-edge AI capabilities with groundbreaking intelligence at the data infrastructure layer to create valuable new solutions. With its evolving smart storage strategy, Google Cloud has taken a meaningful step forward in this regard, with powerful new offerings at the storage level that, when combined with its industry-leading AI capabilities, offer a complete, end-to-end and vertically integrated AI technology stack that is unrivaled in the industry. This, along with Google's vast experience of building and leveraging massive unstructured data repositories of its own, means organizations looking to illuminate their dark data at scale should take a closer look at the smart storage capabilities Google Cloud is developing.

About Enterprise Strategy Group
Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com