Market Forecast

Worldwide Global DataSphere Structured and Unstructured Data Forecast, 2024–2028

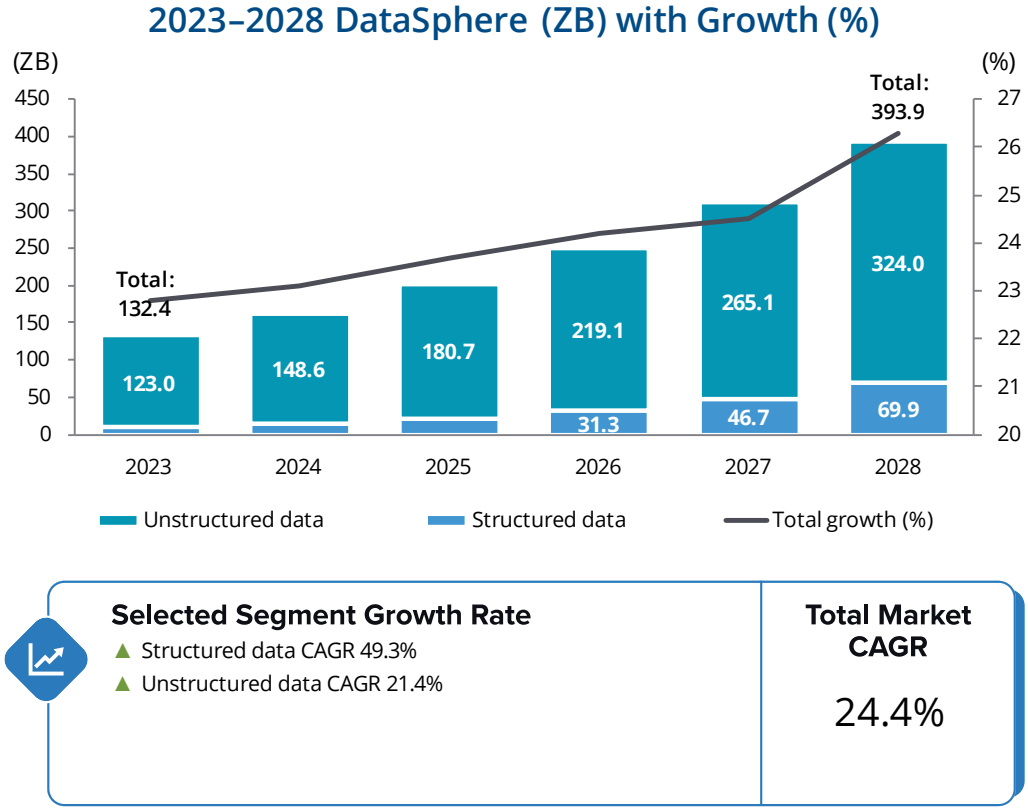
Adam Wright

THIS IDC EXCERPT FEATURES BOX

IDC MARKET FORECAST FIGURE

FIGURE 1

Worldwide Global DataSphere Data Generation Snapshot



Selected Segment Growth Rate

- ▲ Structured data CAGR 49.3%
- ▲ Unstructured data CAGR 21.4%

Total Market CAGR

24.4%

Note: Chart legend should be read from left to right.

Source: IDC, 2024

September 2024, IDC #US52554824e

EXECUTIVE SUMMARY

IDC's Global DataSphere is a measure of how much new data is created, captured, replicated, and consumed each year. It is forecast by several segments including consumer/enterprise, region, data type, location (core, edge, endpoint), and deployment type (cloud/noncloud) of the technology involved in data generation.

IDC further segments IDC's Global DataSphere by the volume of structured and unstructured data generated and stored, which highlights one of the 3D characteristics of the Global DataSphere and Global StorageSphere: data diversity.

In 2023, approximately 92.9% of the data generated was unstructured data, yet over the forecast period, the volume of structured data generated will grow faster than unstructured data generated, resulting in a share of 17.7% in 2028. The high growth rate of structured data over the forecast period is in part driven by the increasing proliferation of IoT devices — which generate a large amount of structured data relative to other data types — as well as by ongoing digital transformation efforts that are implementing various business productivity and intelligence tools that inherently generate and rely on structured data, among other factors.

This IDC study provides a forecast of structured and unstructured data in IDC's Global DataSphere. Various dynamics and trends impacting the mix of structured and unstructured data are highlighted and discussed, and definitions of structured and unstructured data are included in the Market Definition section.

"The mix of structured data versus unstructured data generated each year is beginning to shift more noticeably," says Adam Wright, research manager for IDC's Global DataSphere research program. "This is in part due to the growing impact of generative AI, which holds the potential to amplify the volume and complexity of unstructured data while refining and expanding the generation of structured data. Enterprises will need to adapt their data management and analytics strategies to respond to this changing mix of data to extract more value and deeper insights."

ADVICE FOR TECHNOLOGY SUPPLIERS

- Enhance enterprise intelligence by offering solutions that transform diverse data structures into actionable insights, which will enable organizations to achieve better business outcomes. Continue to invest in AI-driven data management tools by developing and offering solutions that leverage AI to automate the

organization, tagging, and analysis of both structured and unstructured data, which will help clients derive actionable insights more efficiently.

- Evaluate how technology solutions generate, capture, manage, and share metadata across various processes to improve data management and utilization. Understanding and leveraging metadata can simplify the integration of diverse data types and enhance the extraction of actionable insights, particularly from unstructured data sources.
- Develop a framework to help organizations balance the risks and benefits of retaining unstructured data, especially as they explore data intelligence solutions like generative AI. This framework should guide organizations in maximizing the value of their data while minimizing potential security and compliance risks associated with long-term data retention.
- Educate clients on data strategy by offering guidance and educational resources to help them understand the importance of balancing structured and unstructured data in their digital transformation efforts, promoting best practices for data governance and utilization. Ensure that solutions support interoperability across various data formats and systems and enable clients to more easily integrate and utilize structured and unstructured data from diverse sources without compatibility issues.

MARKET FORECAST

IDC's Global DataSphere Structured and Unstructured Data Forecast

The most recent IDC's Global DataSphere forecast is the foundation for calculating the volume of structured and unstructured data in this study (*see Worldwide IDC Global DataSphere Forecast, 2024–2028: AI Everywhere, But Upsurge in Data Will Take Time,* IDC #US52076424, May 2024). Changes to the underlying mix of data (by data type) from year to year in IDC's Global DataSphere forecast can impact the mix and growth rates of structured and unstructured data. Table 1 provides IDC's Global DataSphere structured and unstructured data forecast for 2023–2028. Figure 2 shows the percentage mix for 2018–2028.

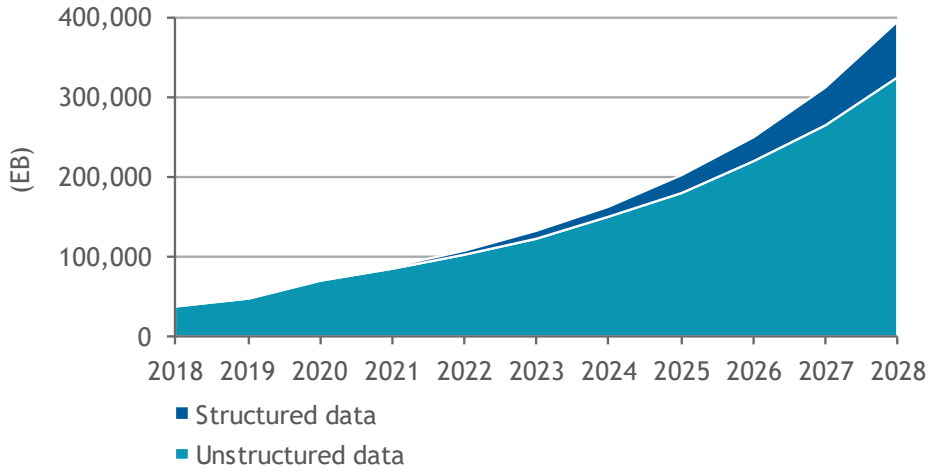
TABLE 1

WORLDWIDE GLOBAL DATASPHERE DATA GENERATION BY STRUCTURED AND UNSTRUCTURED DATA, 2023–2028 (EB)

	2023	2025	2028	2023–2028 CAGR (%)
Structured data	9,433	20,910	69,869	49.3
Growth (%)	47.7	44.9	49.5	
Unstructured data	122,992	180,745	323,983	21.4
Growth (%)	21.2	21.7	22.2	
Total	132,425	201,655	393,852	24.4
Growth (%)	22.8	23.7	26.3	

FIGURE 2

Worldwide Global DataSphere Data Generation by Structured and Unstructured Data, 2018–2028



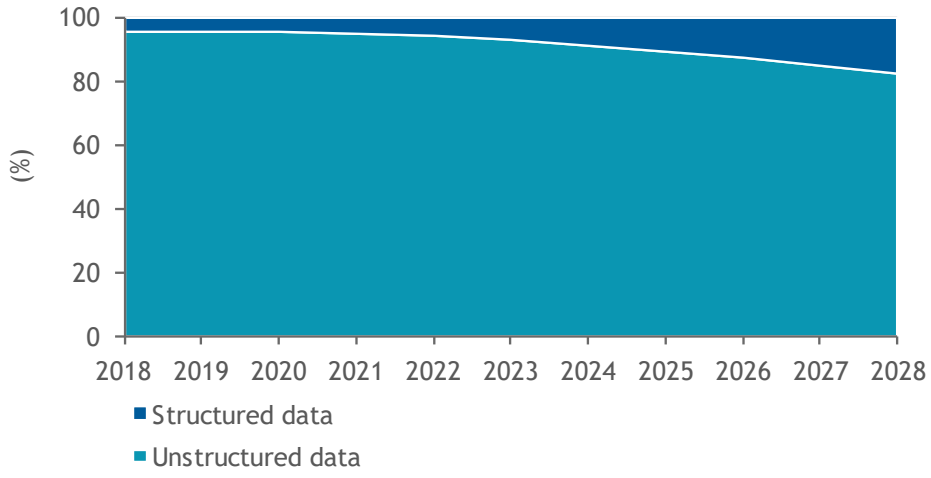
Source: IDC's Global DataSphere, 2024

Like the prior study, IDC expects the volume of structured data generated will grow faster than unstructured data over the forecast period. Unstructured data currently

makes up about 93% of all data generated. Due to faster growth of structured data over the forecast period, the share of unstructured data generated is expected to decline to less than 83% of IDC's Global DataSphere by 2028 (see Figure 3).

FIGURE 3

Worldwide Global DataSphere Data Generation Share by Structured and Unstructured Data, 2018–2028



Source: IDC's Global DataSphere, 2024

Structured Data in IDC's Global DataSphere, Forecast by Data Type

Several specific types of structured data are exhibiting fast growth and are driving the overall volume of structured data growth at a 2023–2028 CAGR of 49.3%. Table 2 shows the composition of structured data generation by data type in IDC's Global DataSphere through 2028.

TABLE 2

Worldwide Global DataSphere Structured Data Generation by Data Type, 2023–2028 (EB)

	2023	2025	2028	2023–2028 CAGR (%)
Entertainment	1	1	1	14.3
Growth (%)	15.0	16.0	14.9	
IoT	337	778	2,960	54.4
Growth (%)	50.3	52.5	58.4	
Non-entertainment image	12	36	139	64.6
Growth (%)	74.0	67.7	51.9	
Productivity	8,955	19,856	66,214	49.2
Growth (%)	47.7	44.7	49.3	
Social media	128	238	555	34.0
Growth (%)	36.6	33.1	32.4	
Voice	0	1	1	6.9
Growth (%)	10.5	7.5	6.1	
Total	9,433	20,910	69,869	49.3
Growth (%)	47.7	44.9	49.5	

The fastest-growing structured data type is non-entertainment image, which represents metadata about objects that are created by using video analytic software that converts video content to objects, interprets each object, and then attaches descriptive metadata about each object. Metadata attached to the video converted to objects provides analytical functionality required to interpret the video data. While non-entertainment image is the fastest-growing structured data type, it is a very small portion of all structured data generated.

IoT structured data is growing at a 2023–2028 CAGR of 54.4% and will make up approximately 4% of structured data generated in 2028. A portion of IoT data is also categorized as streaming data. In 2023, more than 52% of the IoT data created was

streaming IoT data, and the volume of IoT created and captured data is expected to grow as new use cases and applications for IoT sensors are adopted.

In 2023, productivity data makes up roughly 95% of all structured data generated in IDC's Global DataSphere. In recent years, IDC has conducted several surveys asking organizations to provide estimates of the mix of structured and unstructured data generated by their organizations. Based on this primary research (and follow-up discussions with vendors and organizations), the mix of structured and unstructured data generated for some enterprise systems and applications was changed slightly in this forecast iteration (see the Changes from Prior Forecast section), which increased the historical and forecast volume of structured data generated compared with the prior forecast (and, correspondingly, reduced the share of unstructured data generated).

Unstructured Data in IDC's Global DataSphere, Forecast by Data Type

Table 3 shows the composition of unstructured data generation by data type in IDC's Global DataSphere through 2028. Note that data IDC categorizes as semistructured is combined with data IDC categorizes as unstructured.

TABLE 3

Worldwide Global DataSphere Unstructured Data Generation by Data Type, 2023–2028 (EB)

	2023	2025	2028	2023–2028 CAGR (%)
Entertainment	45,302	54,692	70,707	9.3
Growth (%)	13.5	9.2	9.8	
IoT	15,147	36,581	94,644	44.3
Growth (%)	56.5	50.2	34.2	
Non-entertainment image	39,569	49,450	67,428	11.2
Growth (%)	15.4	12.0	10.5	
Productivity	7,579	14,720	43,687	42.0
Growth (%)	36.5	39.3	45.4	
Social media	11,316	20,395	41,604	29.7
Growth (%)	32.9	36.8	24.1	
Voice	4,079	4,909	5,913	7.7
Growth (%)	15.3	8.4	5.7	
Total	122,992	180,745	323,983	21.4
Growth (%)	21.2	21.7	22.2	

The combined volume of entertainment and non-entertainment image data generated represents about 69% of unstructured data generated in IDC's Global DataSphere in 2023 yet is forecast to decline to less than 43% by 2028 because of faster growth of unstructured IoT, social media, and productivity data over the forecast period. The volume of entertainment data generated will continue to grow, but not as fast as a decade ago because of slower adoption of higher-resolution video (e.g., 4K and 8K video). The majority of IoT data is assumed to be semistructured or unstructured, yet only a small percentage of IoT data is saved (as reflected in IDC's Global StorageSphere unstructured data forecast).

Structured and Unstructured Data Generated in the Enterprise DataSphere

IDC's enterprise Global DataSphere also indicates stronger growth for the volume of structured data than unstructured data generated in 2023 and each year over the forecast period (see Table 4). Structured data made up less than 12% of total enterprise data generated in 2023 but will grow to about 22% of enterprise data generated in 2028.

TABLE 4

Worldwide Enterprise Global DataSphere Data Generation by Structured and Unstructured Data, 2023–2028 (EB)

	2023	2025	2028	2023–2028 CAGR (%)
Structured data	9,407	20,857	69,739	49.3
Growth (%)	47.7	44.9	49.5	
Unstructured data	75,469	123,474	247,394	26.8
Growth (%)	27.4	28.4	26.1	
Total	84,876	144,331	317,133	30.2
Growth (%)	29.4	30.5	30.6	

In this forecast iteration, growth in the volume of unstructured data generated by AI (including generative AI) is assumed to continue its trend leading up to 2023. Preliminary research suggests that the growing adoption of generative AI solutions by consumers and enterprises could lead to a content creation surge as organizations and individuals create more data-rich media and documents. In enterprise settings, generative AI is beginning to be leveraged for a range of tasks that could contribute to a steady increase in unstructured data generated, such as automated report generation, customer service interactions, and content marketing, to name a few. Ongoing primary IDC research will assess if faster adoption of generative AI solutions by organizations will accelerate the volume of unstructured data generated and by how much.

Structured data growth at enterprise organizations over the forecast period is being driven mainly by two data types: productivity and IoT data (see Table 5). These two data types make up more than 98% of all structured data generated each year.

TABLE 5

Worldwide Enterprise Global DataSphere Structured Data Generation by Data Type, 2023–2028 (EB)

	2023	2025	2028	2023–2028 CAGR (%)
Entertainment	1	1	1	14.3
Growth (%)	15.0	16.0	14.9	
IoT	315	731	2,838	55.2
Growth (%)	50.2	53.3	59.6	
Non-entertainment image	11	35	138	65.2
Growth (%)	75.6	68.4	52.1	
Productivity	8,952	19,852	66,207	49.2
Growth (%)	47.8	44.7	49.3	
Social media	128	237	553	34.1
Growth (%)	36.7	33.2	32.4	
Voice	0	0	1	7.3
Growth (%)	11.0	7.9	6.5	
Total	9,407	20,857	69,739	49.3
Growth (%)	47.7	44.9	49.5	

Note that both productivity and IoT data types may be structured or unstructured. IDC endeavors to categorize productivity and IoT data as structured and unstructured based on the subtype of data in IDC's Global DataSphere Model. For example, documents created using typical office applications (spreadsheets, documents, presentations) are categorized as unstructured productivity data, whereas enterprise data created or captured by certain business applications (including its replication for analysis) is categorized as structured productivity data.

IoT and productivity data are two of the three fastest-growing data types for enterprise unstructured data (see Table 6). Non-enterprise data historically made up the highest portion of enterprise unstructured data generated (e.g., video surveillance, medical

images, video), but IoT and productivity data combined is expected to make up more than 53% of enterprise unstructured data generated in 2028, up from 28% in 2023.

TABLE 6

Worldwide Enterprise Global DataSphere Unstructured Data Generation by Data Type, 2023–2028 (EB)

	2023	2025	2028	2023–2028 CAGR (%)
Entertainment	8,933	12,694	18,127	15.2
Growth (%)	28.0	17.9	12.6	
IoT	14,569	35,192	90,764	44.2
Growth (%)	56.6	50.2	34.1	
Non-entertainment image	34,307	42,601	57,740	11.0
Growth (%)	14.7	11.7	10.3	
Productivity	6,960	13,962	42,666	43.7
Growth (%)	39.3	41.3	46.5	
Social media	7,549	15,200	33,433	34.7
Growth (%)	41.9	45.2	26.3	
Voice	3,152	3,825	4,664	8.2
Growth (%)	15.9	8.9	6.1	
Total	75,469	123,474	247,394	26.8
Growth (%)	27.4	28.4	26.1	

Drivers and Inhibitors

Drivers

Proliferation of IoT Devices

- **Assumption:** Steady growth of the installed base of connected IoT devices by enterprise organizations is expected to provide a critical source of data for insights, new monitoring capabilities, productivity improvements, cost optimization, and revenue generation. IoT data is also becoming richer, driving higher data traffic on networks.
- **Impact:** The growing installed base of IoT devices and the richness of data created or captured will drive the growth of structured data generated, managed, and analyzed by organizations.

Growth of Data Created in IDC's Global DataSphere

- **Assumption:** The volume of data created each year in IDC's Global DataSphere is forecast to increase at a 2023–2028 CAGR of 25.1%.
- **Impact:** The volume of data stored is expected to grow at a 2023–2028 CAGR of 16.6%, slightly lower than the growth of data created. While not all data created is required to be stored, new abilities to analyze and gain insights from data could drive the growth rate of data stored higher.

Inhibitors

Growing Complexity of a 3D DataSphere and StorageSphere

- **Assumption:** Several factors are adding complexity to modern data management, including the number and variety of data sources, distribution of data across various data silos and/or hybrid and multicloud environments, distribution of data across geographies, variety of data types, and the velocity of data flow or movement. System complexity, and in any system, generally leads to system fragility.
- **Impact:** Growing system fragility could lead to an increase of poor data outputs and data breaches and/or misuse of data that might cause organizations to slow data growth until system complexities (including security vulnerabilities) are solved and simplified.

Near-Term IT Spending Changes

- **Assumption:** Organizations may reduce IT spending plans in 2024 because of volatile economic conditions and company profitability and/or because of higher

IT costs that are being fueled by inflation/currency, IT supply chain disruptions, and labor shortages.

- **Impact:** Historically, growth of data generation in IDC's Global DataSphere has been unimpeded by economic or IT spending downturns. Yet the rate of enterprise data generation growth could be slowed if the IT spending downturn is significant and causes organizations to prioritize defensive (e.g., security) rather than offensive (e.g., improvements to processes and decision-making) projects.

Significant Market Developments

Over the past 18 months, several market developments have significantly impacted the growth of data generation and consumption globally, both in enterprise and in consumer settings. Collectively, these developments are creating a feedback loop in which increased data generation demands more advanced data processing and analysis tools, which in turn spurs further data creation. The ongoing investment in AI, digital content creation, and edge computing is contributing to the increase in global data volumes and will reshape how enterprises and consumers interact with digital technologies. Key highlights include:

- **Expansion of generative AI:** Generative AI has seen significant growth, with major updates to platforms like ChatGPT and companies launching AI-driven products that increase data generation and utilization. ChatGPT expanded its functionalities to include voice and image interactions and also expanded the availability of its app to more regions. Salesforce, Accenture, and others have introduced AI tools that enhance customer interaction and marketing strategies, driving the need for more data collection and analytics. AI's integration in media production also became more apparent, with studios like Paramount and Disney starting to use generative AI tools for tasks ranging from lip syncing to creating complex visual effects and advanced content personalization, to name a few.
- **More digital content creation tools**: Digital content creation tools have proliferated, fueled by the demands of remote work and digital marketing. Companies like Adobe and Microsoft have enhanced their offerings, leading to a surge in data production through graphics, video content, and web applications. This growth is especially pronounced in the Asia/Pacific region, where digital media consumption is booming, contributing to a significant increase in data volume.
- **Advancements in cloud and edge computing:** Over the past 18 months, there has been a steady investment in cloud and edge computing technologies, which has helped facilitate a shift toward decentralizing data processing to edge devices, enabling faster data processing and reduced latency — which is critical

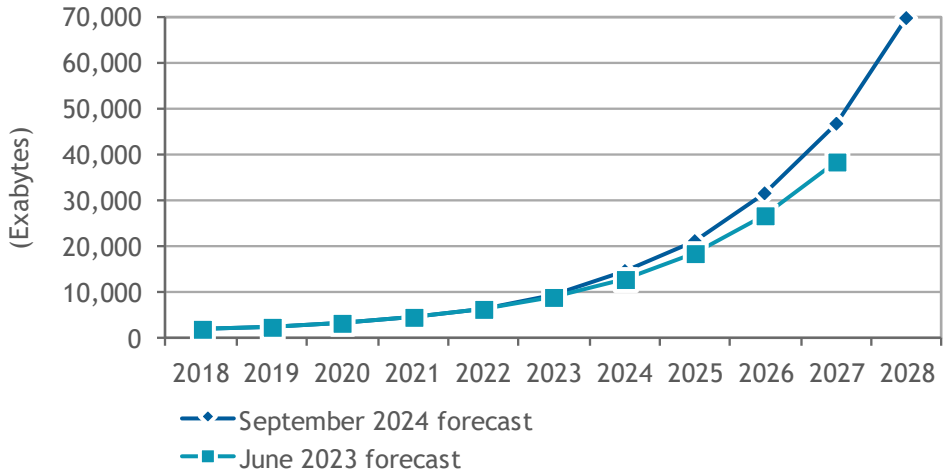
for the expanding IoT and mobile devices market. The adoption of hybrid cloud solutions has also continued to grow, which is helping organizations leverage advantages of both private and public clouds for more flexible, cost-effective, and scalable computing environments. This expansion of edge computing and the IoT is impacting both structured data — such as sensor readings — and unstructured data, like video feeds from cameras or audio inputs from smart devices. This growth is further driven by the need for real-time processing and analysis closer to the source of data generation.

Changes from Prior Forecast

Changes in calculation methodologies and additional diligence applied to data categorization assumptions resulted in a slight shift of unstructured data to structured data compared with the previous forecast (see *Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027,* IDC #US50397723, June 2023). The volume of structured data is expected to grow faster in this year's forecast compared with the prior forecast because the mix of structured and unstructured data generated for some enterprise systems and applications was changed in this forecast iteration. Figures 4 and 5 compare this year's IDC's Global DataSphere structured and unstructured data generated forecast, respectively, with the prior forecast.

FIGURE 4

Worldwide Global DataSphere Structured Data Generation, 2018–2028: Comparison of June 2023 and September 2024 Forecasts

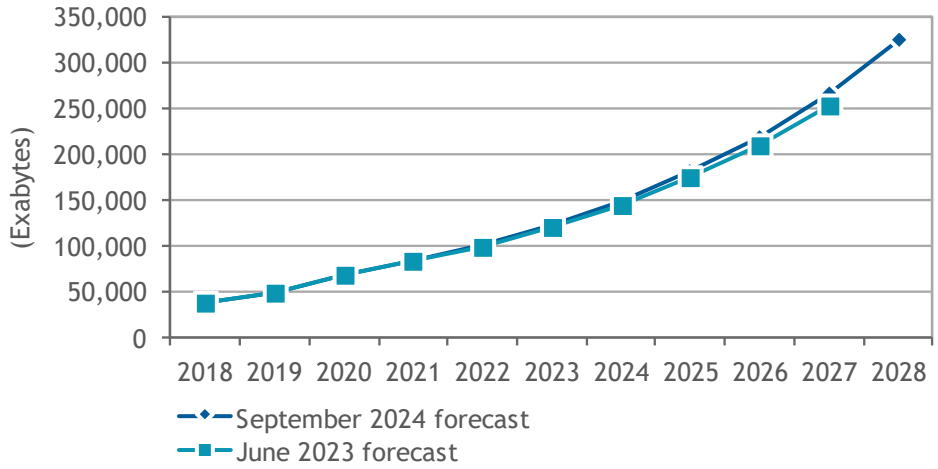


Note: See *Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027* (IDC #US50397723, June 2023) for prior forecast.

Source: IDC's Global DataSphere, 2024

FIGURE 5

Worldwide Global DataSphere Unstructured Data Generation, 2018–2028: Comparison of June 2023 and September 2024 Forecasts



Note: *See Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027* (IDC #US50397723, June 2023) for prior forecast.

Source: IDC's Global DataSphere, 2024

MARKET DEFINITION

IDC's Global DataSphere is a measure of how much new data is created each year. It is not a measure of how much data is stored, which is found in IDC's Global StorageSphere. It is recognized that much more data is created than stored in the Global StorageSphere any given year. Global DataSphere data that is not stored is referred to as ephemeral or cached data.

Data in the Global DataSphere and StorageSphere can be subsegmented and categorized in a variety of ways, including the volume of data generated and/or stored that is structured or unstructured (including semistructured data). The definitions for structured, unstructured, and semistructured data used in this forecast are:

- **Structured:** This represents highly organized data addressable for analysis. It conforms to a data model, is usually stored in relational databases, and is generally tabular with columns and rows (n x n tables). Examples include SQL databases and OLTP systems. Data in a single spreadsheet could be considered structured, but data in multiple spreadsheets is essentially considered to be unstructured files.

- **Unstructured:** This represents data not organized in a predefined manner and that does not conform to a semantic database model that facilitates addressing and/or analyzing the data. It is stored in its native format and may or may not be labeled, tagged, or annotated with contextual metadata. Examples include images, videos (including computer vision telemetry data), text files, and word documents stored in a file system.
- **Semistructured:** This represents data that does not reside in a relational database or conform to a data model but is data that has a structured component for organizational and/or analytical purposes. Typically, the underlying data (or the contents within the data container) itself is unstructured, and a data labeling, tagging, or marker system is used that adds one or more meaningful labels about the data. Alternatively, another self-describing structure is employed to provide context so that a model can learn from the data. Examples include smartphone photos stored in a library, emails stored in an email system, and XML and other markup language data that is organized in a system and web pages.

Data may also be categorized as synthetic as opposed to real, meaning the data is artificially manufactured rather than generated by real-world events, and it may be structured or unstructured. Synthetic data is usually created from existing data sets and is often used instead of real-world data for application testing or AI/ML training, especially because it can be more cost effective and efficient than collecting a sufficient amount of real-world data. IDC believes synthetic data is typically short lived and represents a small portion of the Global DataSphere.

METHODOLOGY

The foundational data sets for the data presented in this study are IDC's Global DataSphere and Global StorageSphere forecasts for 2023–2028. A broad range of sources are used to develop and quantify assumptions for categorizing data as structured or unstructured, including:

- Input from key IDC subject matter experts representing a wide range of research practice areas
- IDC surveys
- Discussions with big data analytics and intelligent knowledge discovery software solution providers
- Discussions with content management solution providers
- IDC server and storage workload data

Note: All numbers in this document may not be exact due to rounding.

RELATED RESEARCH

- *Worldwide Global StorageSphere Forecast, 2024–2028: AI Everywhere, But Storage Capacity Remains a Balancing Act* (IDC #US52312824, June 2024)
- *IDC's Macroeconomic Forecast Assumptions, April 2024* (IDC #US52097924, May 2024)
- *Worldwide IDC Global DataSphere Forecast, 2024–2028: AI Everywhere, But Upsurge in Data Will Take Time* (IDC #US52076424, May 2024)
- *IDC's Worldwide Global StorageSphere Taxonomy, 2024* (IDC #US50614024, March 2024)
- *IDC's Worldwide Global DataSphere Taxonomy, 2024* (IDC #US50613924, March 2024)
- *World Backup Day: Highlighting the Importance of Data Resilience* (IDC #lcUS51996624, March 2024)
- *Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027* (IDC #US50397723, June 2023)

ABOUT IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com