SOLIX®
Empowering the Data-driven Enterprise

THE REINVENTION OF DATA:

TRANSFORMING YOUR FORGOTTEN DATA INTO AI INTELLIGENCE

info@solix.com

https://www.solix.com

+1 408-654-6400

JOHN OTTMAN
Executive Chairman
Solix Technologies, Inc.

Spring 2025

The reinvention of data

Data never lost its value — we just stopped paying attention. Now, with Enterprise AI, we can finally unlock that value — especially in unstructured formats that were once too complex and too costly to decode. The problem for most organizations is that so much unstructured data is largely inaccessible and useless. At a time when generative AI systems require the highest quality enterprise data to provide business context for accurate prompt responses, unstructured data remains a significant untapped asset

Unstructured data is inconsistent and lacks a predefined data model or schema, making it difficult to process and utilize effectively. The data is mostly text files, reports and PDFs, but also includes spreadsheets, images, audio and video files. Unlike structured data which is organized nicely into tables and databases with rich metadata, unstructured data is often unclassified, uncatalogued, and is stored in S3 buckets as raw files or in silos of network drives spread out across the enterprise.

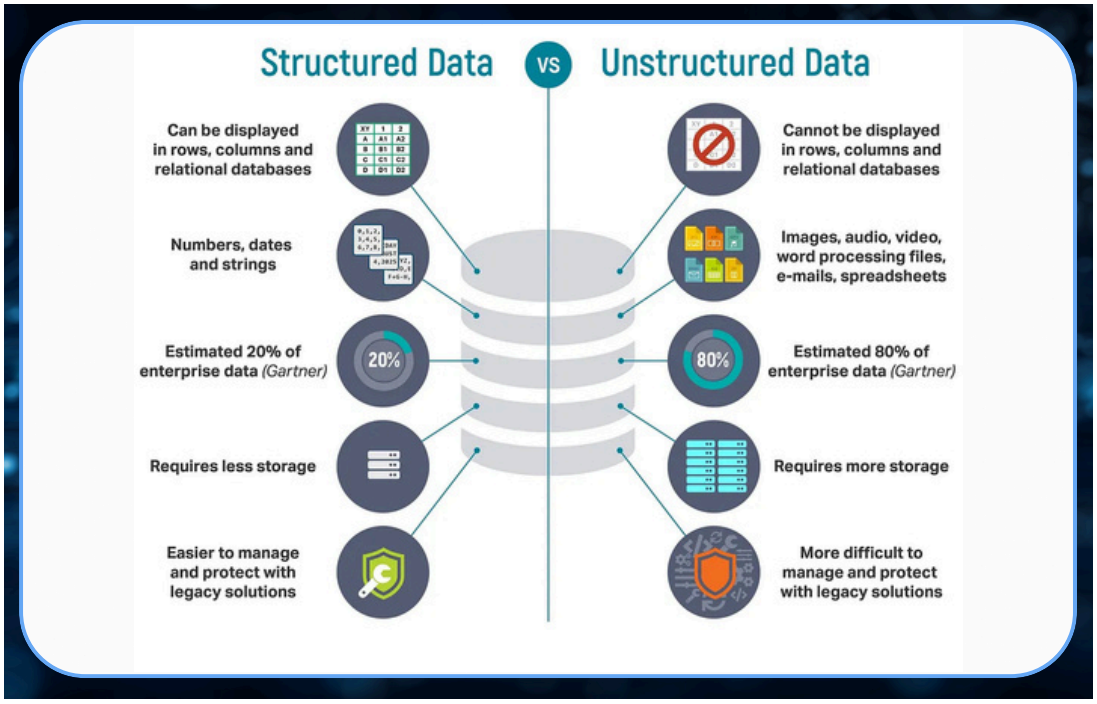

Image source: https://www.igneous.io/blog/structured-data-vs-unstructured-data

How much data?

The scale of unstructured data is enormous, and the rate of growth is exponential. Unstructured data accounts for up to 80 percent of all enterprise data and is growing at an astounding 55% to 65% per year. The sources of data are vast and numerous and include server logs, Internet of Things (IoT) sensors, document stores, customer systems, social media, chat systems and emails.

In 2023, there were 347 BILLION emails sent and received globally. Knowing that emails

IDC Report "The Digitization of the World From Edge to Core [(1)(2)]

are still the backbone of enterprise business communication, this is a gigantic figure. Think about what the accumulation of all this email data means for management and governance, especially when so many emails come with attachments (aka unstructured data) that get saved onto corporate drives. With semi-structured and unstructured data growing three times faster than structured data, enterprises face complex data governance and data management problems.

The origins of dark data

According to research by True Global Intelligence and Splunk, 60% of surveyed respondents reported that half or more of their organization's data is considered "dark" — unquantified and untapped — and a full one-third of respondents reported this amount to be 75% or more. While the 1,300 business leaders surveyed recognized the importance of data to their success, few were able to report that their organizations could successfully tap the value of their dark data, or in some cases, even find it.

Over time huge stores of unstructured data are created, but the data is seldom used for anything at all, let alone business intelligence or decision-making. Lacking proper data governance and Information Lifecycle Management (ILM) controls, most organizations regard this aging data as essentially unusable by anyone other than the person who created it, assuming that person still exists within the organization.

Lacking clear insights into the contents of the files, the data may not be safely cataloged, governed or deleted, so instead it piles up over time. For medium to large size organizations these stores of uncatalogued, unclaimed dark data represent a petabyte scale compliance concern.

The impact of dark data

Enterprise data becomes less active as it ages, and by eighteen months it becomes nearly inactive. Despite its AI value, the data remains untapped. But perception is reality: In the real-time world where data freshness is critical to data quality, users will quickly discount stale data as untrustworthy, leaving it to become dark and forgotten. The worst of it becomes ROT — Redundant, Obsolete and Trivial — which includes copies and different versions of the same information, outdated information, information that is no longer in use, and extraneous or irrelevant information that is of no value.

Throughout the data lifecycle systems are upgraded, archived and retired, and legacy data often goes dark as soon as it goes offline. Distributed or isolated user groups may collect and store their own data locally, and these datasets may become stranded and lost because they

are not observable centrally. And, of course, sometimes a change in business strategy leads data to going dark because no one cares about that data anymore. However it happens, once data goes dark it is no longer usable by others who might actually find it valuable. (4)

Impact Area	Effect of Dark Data
Cost	*Higher storage, backup, and management costs*
Security and Compliance	*Risk of breaches, regulatory fines, and reputation damage*
Operational Efficiency	*System bloat, data overload, reduced agility*
Environmental	*Increased power usage and carbon footprint*
Innovation & Insight	*Missed opportunities for analytics, automation, and value creation*

What are the risks?

In many organizations unstructured data lacks a data governance strategy such as Information Lifecycle Management (ILM) which has been used successfully with structured data for so many years to provide policy based governance controls. In the absence of a proper data governance framework, compliance failures should be expected.

At a minimum data governance controls start with consistent and accurate data classification to determine data security, access control and regulatory compliance requirements. The data classification process may be manual, automated, or a hybrid of both, as data quality may require human-in-the-loop oversight regardless.

Legal hold and data retention policies are more difficult regarding large stores of unstructured and-semi structured data because the contents of the files may not be known. The threat that these unstructured files may contain sensitive or personally identifiable information (PII) is high, and controls are required or security teams may face no alternative other than to lock down access.

When unstructured data contains sensitive and confidential information such as credit card or personal healthcare records, the same rigorous security and compliance controls commonly applied to structured data are required, but these measures are rarely if ever implemented. As a result unstructured data often lacks permissioned access control, and should not be shared across the enterprise

Dark, unstructured data poses a broad array of compliance and marketplace risks including fines, customer dissatisfaction, brand damage and/or other liabilities which can occur if an unplanned security or compliance incident were to occur. Without compensating controls to manage these risks, organizations are vulnerable to violations of consumer data privacy laws and compliance frameworks such as FISMA, HIPAA and PCI-DSS.

Risk Category	Description
Data Quality	*Inaccurate or inconsistent data undermines analytics & operations*
Compliance & Legal	*Noncompliance with data regulations leads to fines and lawsuits*
Security	*Unprotected data is vulnerable to breaches and leaks*
Ownership & Stewardship	*No clear roles lead to chaos and ungoverned data silos*
Shadow IT	*Unapproved tools introduce risk and fragmentation*
Business Intelligence	*Untrustworthy data leads to poor insights and decisions*
Traceability	*No audit trail means no accountability*

What are the costs?

An S3 bucket costs just $0.023 per gigabyte per month in AWS US East, proving that data storage is indeed cheap, but other factors also influence the cost of storing massive amounts of dark, unstructured data. For instance:

- Larger and larger datasets often carry a super-size effect on the cost of other infrastructure components including servers, memory and networks.
- The cost of reputational damage emerges when dark data containing sensitive information is improperly accessed or inadvertently exposed. Furthermore, the cost of not being able to respond to regulatory audits or consumer data privacy requests may result in legal and professional service fees, and even fines.
- And let's not forget that employees spend lots of time searching for relevant information, and ungoverned, uncatalogued data reduces productivity and increases labor costs as workers struggle to find what they are looking for. The highest cost of all being when workers completely fail to find the information they need, and business outcomes suffer as a result.

Enterprise AI impact

Enterprise AI has arrived with a bang, and in a sudden and serendipitous series of events, unstructured data has taken center stage. Just as the surrender flag was about to be raised over how to manage the compliance risk of unstructured data, or if it would ever create value, a gold rush to collect, preserve and repurpose unstructured data has begun. Overnight, these same petabytes of forgotten, unstructured data have transformed into a priceless asset critical for generative AI and autonomous, agentic workflows delivering value across the data-driven enterprise.

As Splunk's research remarks, "dark, unstructured data may be an organization's biggest untapped resource." Business leaders are now

moving rapidly to capture the enterprise intelligence stored in their unstructured data. Within these petabytes of unknown, unclaimed data lies the most complete history of the enterprise that exists complete with email trails, copies of payments, invoices, etc..

Getting started

Not surprisingly, the best way to reclaim unstructured data is with AI. Intelligent Data Classification (IDC) reads text, image and even video files, and then automatically abels sensitive, confidential or personally identifiable information (PII). Once the files are read and classified, rich metadata may be created enabling data governance rules and Role Based Access Control (RBAC). The ability for AI to automate the classification of a large corpus of dark, unstructured data is a game changer that simply was never possible before.[(4)]

AI safety and security depends on robust data governance controls. With the data properly classified, security and compliance controls may be used to protect sensitive data from unauthorized access and to manage data retention and legal hold policies.

Data unification is another important goal realized through an AI data warehouse strategy where local data, indexes and metadata are stored. The goal is a single catalog view of all AI data with no data copy or requirement to move any data. By sharing open metadata, federated data governance controls may be deployed regardless of the physical location of the data.

Data preparation for AI is a critical step, and pipelines must be built and maintained to collect data from any source, and deliver it in real-time to

Enterprise AI for processing. AI semantics simplifies data access, improves data consistency and accuracy, and enhances the performance and reliability of AI applications. By providing a unified, business-friendly view of data, AI semantics improves accuracy, limits hallucinations, and reduces unnecessary inference processing.

Reinvention of data rewards

Enterprise AI has arrived with a promise to raise enterprise productivity and performance to new levels. But first we must activate dark, unstructured data and combine it with existing structured data assets to create new and enriched datasets optimized for AI agents and advanced analytics. And the entire process needs to happen at an incredibly low and affordable cost. The success of autonomous agents traversing the enterprise and executing superhuman tasks depends on the highest quality enterprise data underneath.

IDC and RBAC are effective controls to classify, govern and protect all types of sensitive information whether structured or unstructured. RBAC grants access to sensitive data based on the roles and responsibilities of

the user, and by masking or redacting sensitive contents of the file such as a credit card or social security number, or by blocking access to the file entirely.

Enterprise AI requires high quality data to deliver accuracy, safety, security and performance as well as huge business benefits including:

- Deeper customer insights by understanding feedback for personalized products and services
- Smarter decision making that combines structured metrics with qualitative trends to uncover growth opportunities
- Higher productivity searching and retrieving information
- Improved data governance across the enterprise
- Real-time data quality

While the impact has only just begun to be felt, we can be certain that the success of Enterprise AI rests on the availability of massive amounts of prepared data to provide deep, business context for accurate, safe and secure responses. For some organizations 'The Reinvention of Data' has already begun as chatbots, agentic workflows and retrieval augmented generation (RAG) systems are now being deployed with live enterprise data. For others the journey of 'Transforming Your Forgotten Data Into AI Intelligence' has just begun. Either way, the first step is to ensure your data is AI-ready!

About Solix Technologies

Solix Technologies, Inc. is a leading provider of enterprise data, AI and data fabric solutions and is trusted by Fortune 2000 companies for digital transformation and data-driven operations. The Solix Common Data Platform (CDP) is a cloud native, enterprise data platform for cloud data management applications including Enterprise Data Lake, Enterprise Archiving, Enterprise Security and Compliance and Enterprise AI. Solix is headquartered in Santa Clara, California, and operates worldwide through direct sales and an established network of value-added resellers (VARs) and systems integrators.

Connect with Solix:

- Visit our website: https://www.solix.com
- To schedule a demo click here
- To discover more and explore these capabilities, visit https://solix.com/products/enterprise-ai.
- Additionally, click here to download our SOLIXCloud Enterprise Data Lake whitepaper.

Acknowledgements:

I am grateful to the following individuals for their feedback and support in the preparation of this paper.

Sai Gundavelli : Founder & CEO, Solix Technologies, Inc.
George Hall : Principal, Credit Breakpoint Advisors Global
Dr. Joseph Lancaster : VP Enterprise AI, Solix Technologies, Inc.
Suresh Mani : Chief Architect - Enterprise AI, Solix Technologies, Inc.
Dr. James Short : Director, SPARK AI Consortium, San Diego Supercomputer Center, UCSD
Robert Dayton : Director Enterprise AI, Solix Technologies, Inc.
Stephen Tallant : Director Product Marketing, Solix Technologies, Inc.
Dr. Richard Wang : Director, Chief Data Officer and Information Quality, (CDOIQ)

Footnotes:

1 IDC, IDC Report "The Digitization of the World From Edge to Core"
2 Statista, https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/
3 Splunk, Inc., https://www.splunk.com/en_us/form/the-state-of-dark-data.html
4 IBM,Inc.,https://www.ibm.com/think/topics/darkdata#:~:text=In%20Splunk's%20global%20research%20survey,be%2075%20percent%20or%20more.
5 Spunk, Inc., https://www.splunk.com/en_us/form/the-state-of-dark-data.html

Bibliography

1 DeepTalk, https://www.igneous.io/blog/structured-data-vs-unstructured-data via https://deep-talk.medium.com/80-of-the-worlds-data-is-unstructured-7278e2ba6b73
2 IDC, IDC Report "The Digitization of the World From Edge to Core"
3 Statista, https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/
4 Splunk, Inc., https://www.splunk.com/en_us/form/the-state-of-dark-data.html
5 IBM,Inc.,https://www.ibm.com/think/topics/darkdata#:~:text=In%20Splunk's%20global%20research%20survey,be%2075%20percent%20or%20more.
6 Breakpoint Advisors Global, Inc., https://breakpointadvisors.global/

SOLIX™
Empowering the Data-driven Enterprise

Stay Connected

www.solix.com info@solix.com +1.888.467.6549