FEATURE

Unlocking the hidden value of dark data

The chances are that most of the data you collect — from human communications to machine logs — is piling up with little plan for actualizing its potential. Good governance and AI can help.

By Maria Korolov

Contributing writer, CIO | AUG 11, 2022 3:00 AM PDT

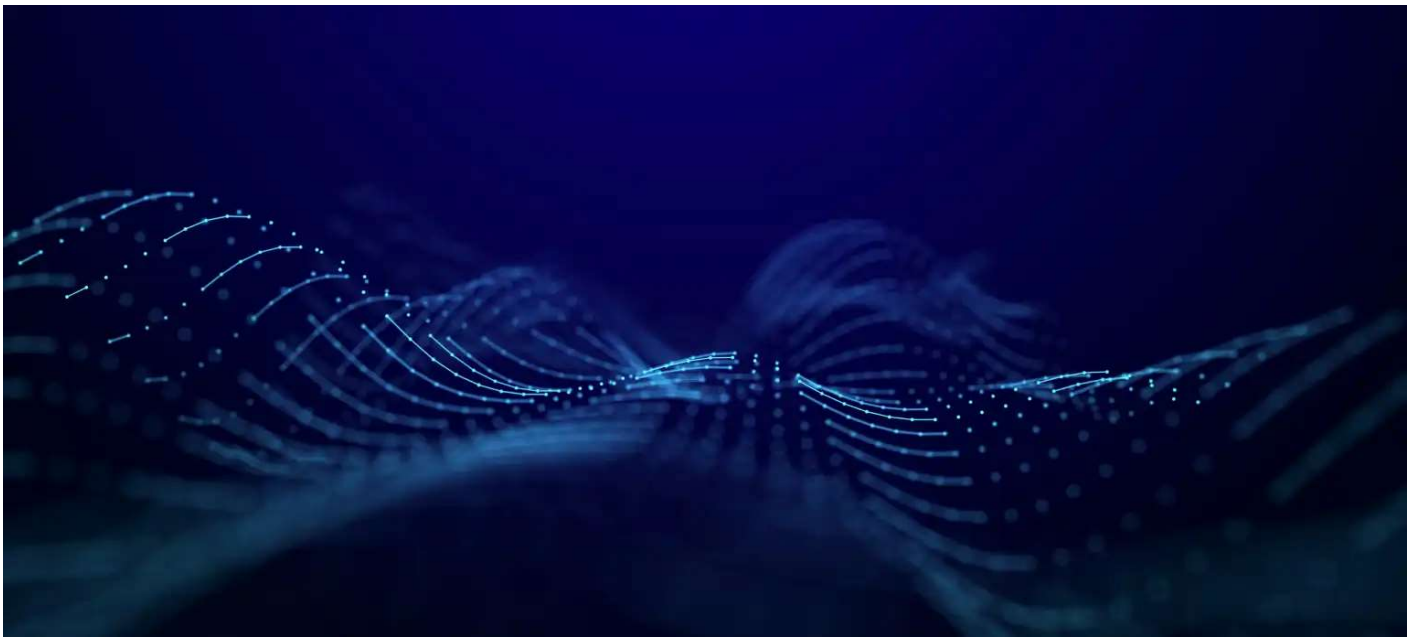

Image: Shutterstock / Olena.07

IT leaders seeking to derive business value from the data their companies collect face myriad challenges. Perhaps the least understood is the lost opportunity of not

making good on data that is created, and often stored, but seldom otherwise interacted with.

This so-called "dark data," named after the dark matter of physics, is information routinely collected in the course of doing business: It's generated by employees, customers, and business processes. It's generated as log files by machines, applications, and security systems. It's documents that must be saved for compliance purposes, and sensitive data that should never be saved, but still is.

[Learn the basics of master data management and discover which MDM certifications will give your career a boost. | Sign up for CIO newsletters.]

According to Gartner, the majority of your enterprise information universe is composed of "dark data," and many companies don't even know how much of this data they have. Storing it increases compliance and cybersecurity risks, and, of course, doing so also increases costs.

Figuring out what dark data you have, where it is kept, and what information is in it is an essential step to ensuring the valuable parts of this dark data are secure, and those that shouldn't be kept are deleted. But the real advantage to unearthing these hidden pockets of data may be in putting it to use to actually benefit the business.

But mining dark data is no easy task. It comes in a wide variety of formats, can be completely unformatted, locked away in scanned documents or audio or video files, for example.

Here is a look at how some organizations are transforming dark data into business opportunities, and what advice industry insiders have for IT leaders looking to leverage dark data.

Coded audio from race car drivers

For five years, Envision Racing has been collecting audio recordings from more than 100 Formula E races, each with more than 20 drivers.

"The radio streams are available on open frequencies for anyone to listen to," says Amaresh Tripathy, global leader of analytics at Genpact, a consulting company that helped Envision Racing make use of this data.

Previously the UK-based racing team's race engineers tried to use these audio transmissions in real-time during races, but the code names and acronyms drivers used made it difficult to figure out what was being said and how it could be made use of, as understanding what other drivers were saying could help Envision Racing's drivers with their racing strategy, Tripathy says.

"Such as when to use the attack mode. When to overtake a driver. When to apply brakes," he says.

Envision Racing was also collecting sensor data from its own cars, such as from tires, batteries, and breaks, and purchasing external data from vendors, such as wind speed and precipitation.

Genpact and Envision Racing worked together to unlock the value of these data streams, making use of natural language processing to build deep learning models to analyze them. The process took six months, from preparing the data pipeline, to ingesting the data, to filtering out noise, to deriving meaningful conversations.

Tripathy says humans take five to ten seconds to figure out what they're listening to, a delay that made the radio communications irrelevant. Now, thanks to the AI model's predictions and insights, they can now respond in one to two seconds.

In July, at the ABB FIA Formula E World Championship in New York, the Envision Racing team took first and third places, a result Tripathy credits to making use of what was previously dark data.

Dark data gold: Human-generated data

Envision Racing's audio files are an example of dark data generated by humans, intended for consumption by other humans — not by machines. This kind of dark data can be extremely useful for enterprises, says Kon Leong, co-founder and CEO of ZL Technologies, a data archiving platform provider.

"It is incredibly powerful for understanding every element of the human side of the enterprise, including culture, performance, influence, expertise, and engagement," he says. "Employees share absolutely massive amounts of digital information and knowledge every single day, yet to this point it's been largely untapped."

The information contained in emails, messages, and files can help organizations derive insights such as who are the most influential people are in the organization. "Eighty percent of company time is spent communicating. Yet analytics often deals with data that only reflects 1% of our time spent," Leong says.

Processing human-generated unstructured data is uniquely challenging. Data warehouses aren't typically set up to handle these communications, for example. Moreover, collecting these communications can create new issues for companies to deal with, having to do with compliance, privacy, and legal discovery.

"These governance capabilities are not present in today's concept of a data lake, and in fact by collecting data into a data lake, you create another silo which increases privacy and compliance risks," Leong says.

Instead companies can also leave this data where it currently resides, simply adding a layer of indexing and metadata for searchability. Leaving the data in place will also keep it within existing compliance structures, he says.

Effective governance is key

Another approach to handling dark data of questionable value and origin is to start with traceability.

"It's a positive development in the industry that dark data is now recognized as an untapped resource that can be leveraged," says Andy Petrella, author of *Fundamentals of Data Observability*, currently available in pre-release form from O'Reilly. Petrella is also the founder of data observability provider Kensu.

"The challenge with utilizing dark data is the low levels of confidence in it," he says, in particular around where and how the data is collected. "Observability can make data lineage transparent, hence traceable. Traceability enables data quality checks that lead to confidence in employing these data to either train AI models or act on the intelligence that it brings."

Chuck Soha, managing director at StoneTurn, a global advisory firm specializing in regulatory, risk, and compliance issues, agrees that the common approach to tackling dark data — throwing everything into a data lake — poses significant risks.

This is particularly true in the financial services industry, he says, where companies have been sending data into data lakes for years. "In a typical enterprise, the IT department dumps all available data at their disposal into one place with some basic metadata and creates processes to share with business teams," he says.

That works for business teams that have the requisite analytics talent in-house or that bring in external consultants for specific use cases. But for the most part these initiatives are only partially successful, Soha says.

"CIOs transformed from not knowing what they don't know to knowing what they don't know," he says.

Instead, companies should begin with data governance to understand what data there is and what issues it might have, data quality chief among them.

"Stakeholders can decide whether to clean it up and standardize it, or just start over with better information management practices," Soha says, adding that investing in extracting insights from data that contains inconsistent or conflicting information would be a mistake.

Soha also advises connecting the dots between good operational data already available inside individual business units. Figuring out these relationships can create rapid and useful insights that might not require looking at any dark data right away, he says. "And it might also identify gaps that could prioritize where in the dark data to start to look to fill those gaps in."

Finally, he says, AI can be very useful in helping make sense of the unstructured data that remains. "By using machine learning and AI techniques, humans can look at as little as 1% of dark data and classify its relevancy," he says. "Then a reinforcement learning model can quickly produce relevancy scores for the remaining data to prioritize which data to look at more closely."

Using AI to extract value

Common AI-powered solutions for processing dark data include Amazon's Textract, Microsoft's Azure Cognitive Services, and IBM's Datacap, as well as Google's Cloud Vision, Document, AutoML, and NLP APIs.

In Genpact's partnership with Envision Racing, Genpact coded the machine learning algorithms in-house, Tripathy says. This required knowledge of Docker, Kubernetes, Java, and Python, as well as NLP, deep learning, and machine learning algorithm development, he says, adding that an MLOps architect managed the complete process.

Unfortunately, these skills are hard to come by. In a report released last fall by Splunk, only 10% to 15% of more than 1,300 IT and business decision makers surveyed said their organizations are using AI to solve the dark data problem. Lack of necessary skills was a chief obstacle to making use of dark data, second only to the volume of the data itself.

A problem (and opportunity) on the rise

In the meantime, dark data remains a mounting trove of risk — and opportunity. Estimates of the portion of enterprise data that is dark vary from 40% to 90%, depending on industry.

According to a July report from Enterprise Strategy Group, and sponsored by Quest, 47% of all data is dark data, on average, with a fifth of respondents saying more than 70% of their data is dark data. Splunk's survey showed similar findings, with 55% of all enterprise data, on average, being dark data, and a third of respondents saying that 75% or more of their organization's data is dark.

And the situation is likely to get worse before it gets better, as 60% of respondents say that more than half of the data in their organization is not captured at all and much of it is not even understood to exist. As that data is found and stored, the amount of dark data is going to continue to go up.

It's high time CIOs put together a plan on how to deal with it — with an eye toward making the most of any dark data that shows promise in creating new value for the business.

Next read this

- *10 IT resolutions for 2022*
- *11 lies CIOs will tell themselves in 2022*
- *IT leaders' top 15 takeaways from 2021*
- *7 hot IT budget investments — and 4 going cold*
- *7 enterprise architecture mistakes to avoid*
- *13 most difficult-to-fill IT jobs*
- *7 hot digital transformation trends — and 3 going cold*
- *7 IT metrics that matter most*
- *7 toxic team behaviors IT leaders must root out*
- *10 key skills for a successful cloud strategy*

Author: Maria Korolov, Contributing writer

Recent stories by Maria Korolov:

- Data lakehouses give enterprises analytics edge
- 6 business risks of shortchanging AI ethics and governance
- 10 tips for getting started with decision intelligence

Top 7 challenges IT leaders will face in 2022

The voice of IT leadership

POLICIES

ABOUT

MORE FROM CIO

DIGITAL MAGAZINE

TOP OF PAGE