IBM

developerWorks®

How data becomes knowledge, Part 3: Extracting dark data

Vinay R. Rao

March 08, 2018

Individuals and organizations store all kinds of data. What do we do with it all? Can we call it up as we need it? Can all that data be analyzed quickly and efficiently? Or, does it tie up storage resources and languish for years because the cost of going through it and discarding what's obsolete is too high? Discover the utility and wisdom of storing dark data.

In my previous article, you saw how data lakes help speed up and reduce the costs of data ingestion by allowing storage of large volumes of multiformat data. The advent of inexpensive storage technologies makes it easier for organizations to store enormous amounts of data cheaply.

Organizations store data for many reasons, most often for record keeping and regulatory compliance. But, there is also a tendency to hoard data that could potentially become valuable. In the end, most companies never use even a fraction of the data they store for any purpose because the data may become inaccessible. This could be because the storage reservoir doesn't document the metadata labels appropriately, some of the data is in a format the integrated tools can't read, or the data isn't retrievable through a query. (This last is especially true for data such as scanned documents, medical imaging files, voice recordings, video, and some forms of machine-generated data.)

This untapped data that organizations routinely store during normal operations is called *dark data.* Dark data is a major limiting factor in producing good data analysis because the quality of any data analysis depends on the body of information accessible to the analytics tools, both promptly and in full detail.

> Gartner defines *dark data* as "the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes."

Companies aren't the only ones who deal with dark data: Many everyday examples of dark data exist. For example, I read a lot of technical papers and journals; often during my research, I download and store PDF files or links for later reference. These files don't have descriptive names and many, especially research papers, simply use a numeric document identifier. Without descriptive information, it becomes impossible to search for a specific article by keywords. To find a particular paper, I might need to open and review each document until I get to the one I want — a time-consuming and inefficient process. And, I often wind up performing the online search all over again only to realize that I already have the file when a download attempt results in a duplicate file error.

I could have mitigated this problem through better data governance, such as storing the files in folders by category or adding descriptive metadata to the file properties. However, doing this consumes time during my search and distracts my train of thought. The result is that I wind up with a collection of often duplicate files that I might never actually use but hoard because they might become useful in the future. In other words, my Downloads folder — my personal data lake — has turned into a data swamp.

Another example of everyday dark data occurs with digital photography. Digital cameras usually follow a file-naming convention of numbering picture files sequentially, and the programs that download images to a computer drive or the cloud typically have a date-based organization. However, if you want to search for photographs of a specific location, person, or event, you have to manually review the photographs because no documentation of the correlation between the photograph creation date the context of the search exists. Photographs embed metadata, but only professional photographers tend to use this feature.

Smart applications have solved both of these problems, initially by using rules-based search and sort methods, but increasingly by using machine learning and deep learning. Desktop search tools can scan through document contents and find documents based on keywords, and photo-organizing tools can recognize faces, landmarks, and features to categorize photographs automatically.

This installment of the series discusses the factors that lead to the creation of dark data, the steps you can take to curate and manage data more effectively, and the methods you can use to extract and use dark data after the fact.

Why does data go dark?

Data becomes inaccessible and unusable for many reasons, but the principal reason is that big data is, well, *big.* Not just big, but mind-bogglingly enormous. Look at a few social media statistics: In 2017, every minute on average, Twitter users sent a half million tweets, and 4 million Facebook users clicked Like.

The 3 Vs of big data

- **Volume**: Big data typically has an enormous volume, and processing this data is both costly and time-consuming. That's why organizations tend to defer processing until doing so is necessary and justifiable. For example, US federal mandates to use electronic medical records forced healthcare organizations to digitize their paper records. But most of these records are in the form of scanned images. A doctor can easily pull up a patient record, but the data within that record isn't accessible to an information retrieval and analysis system.
- **Variety**: Data also comes in a large variety of formats, both structured and unstructured. For example, customer relationship management (CRM) data typically includes email messages, social media messages, voice messages, video, and so on in addition to traditional data

in a database. Formats such as audio, images, and video need preprocessing to extract information for storage in a format conducive to retrieval through query and analysis. Again, for reasons of cost and time, organizations tend to defer this preprocessing and simply store the raw data.

- **Velocity**: Business transactional and operations systems such as stock market trades or card transactions in the financial industry can generate high-velocity data streams. The processing and structuring of such data often lags behind the data arrival rate. An organization often stores this data just for regulatory compliance and auditing. Because there's no immediate need to process the data, the result is to defer processing in favor of storing raw data.

Lack of data provenance

In this case data is accessible but has no provenance. It is simply not usable for analysis. The raw unstructured data is needed for provenance, but it is not accessible. Result: dark data.

This is not a direct relationship. Data scientists rely on the credibility and trustworthiness of data sources to ensure that the product of data analysis is credible and reproducible. If data doesn't have a provenance, then it becomes unusable as a reliable source of information. Part 2 showed that data lakes facilitate curating this provenance by preserving the unstructured and raw data.

Poor metadata documentation

Another common reason for a data source to become unusable is the lack of good metadata. Missing metadata leads directly to data becoming dark data because you cannot access the data through queries. Inferior quality or incorrect metadata also causes good data to become inaccessible through metadata searches. Similarly, inconsistent metadata can split a category based on the variations in the label metadata.

The pitfalls and risks of dark data

Now that you have seen how data turns into dark data, it is time to examine the pitfalls and risks associated with dark data.

Data quality

The main impact of dark data is on the quality of data used for analysis to extract valuable information. This is important. Dark data makes it difficult to access and find vital information, confirm its origins, and promptly obtain essential information to make good, data-driven decisions. The impact on quality stems from the following factors:

- **Data accessibility**: Inability to access data that is unstructured or in a different media format, such as images, audio, or video, leads to loss of access to essential information that would improve analysis.
- **Data accuracy**: The accuracy of a data analysis rests on the accuracy of the input data. Accurate analysis leads to the extraction of qualitatively more valuable information. Hence, dark data has a significant impact on the accuracy of the extracted information and the quality of the information produced by that analysis.

- **Data auditability**: The inability to trace the provenance of data can lead to its omission from analysis, thereby affecting the data quality. This, in turn, can lead to faulty data-driven decision making.

Data security

Stored data often holds sensitive information. Sensitive information could include proprietary information, trade secrets, personal information of employees and clients such as financial and medical records, and so on. Organizations tend to relax data security processes when they do not know that their data store holds sensitive information. Data security breaches are on the rise from hackers who often discover this sensitive information first. This leads to costly liability and remedial actions.

Increased costs

Dark data leads to higher costs in two ways:

- **Data storage costs**: Although data storage hardware costs are decreasing, the volume of stored information grows exponentially and can add up significantly in the long term. With third-party storage management solutions, the result is the application of a higher subscription tier, which in turn leads to spiraling costs. This added cost is for data that has unknown worth as it is dark data.
- **Regulatory compliance**: Businesses must follow many laws and regulations. Some, such as the Sarbanes-Oxley Act, drive the need to store business-related data; others, such as the Health Insurance Portability and Accountability Act and Payment Card Industry Data Security Standard, have requirements for enhanced protection of certain sensitive stored data, all of which can lead to increased compliance monitoring costs. Organizations also incur an added cost for monitoring and securely destroying expired data. As a result, organizations might continue to store dark data long after the regulatory period has lapsed as both the sensitivity details or whether the data has expired are unknown.

The benefits of extracting dark data

Organizations extracting dark data incur an expense and spend considerable engineering effort, but there are many benefits to doing this.

Dark data is valuable

Dark data is valuable because it often holds information that is not available in any other format. Therefore, organizations continue to pay the cost of collecting and storing dark data for compliance purposes and with hopes of exploiting the data (for that valuable information) in the future.

Because of this value, organizations sometimes resort to human resources to manually extract and annotate the data, and then enter it into a relational database, even though this process is expensive, slow, and error-prone. Deep learning technologies perform dark data extraction faster and with much better accuracy than human beings. Dark data extraction is less expensive and uses less engineering effort when using these techniques and tools.

Better-quality analytics

With access to better data sources and more information, the quality of analytics improves dramatically. Not only is the analysis based on a larger pool of high-quality data, but the data is available for analysis promptly. The result is faster and better data-driven decision making, which in turn leads to business and operational success.

Reduced costs and risks

Extracting dark data leaves organizations less exposed to risks and liability in securing sensitive information. Organizations can also securely purge unnecessary data, thereby reducing the recurring storage and curation costs. Regulatory compliance also becomes easier.

Dark data extraction technology is valuable

In addition to the dark data itself, dark data extraction technologies are extremely valuable. Recent reports suggest that Apple purchased artificial intelligence (AI) company Lattice Data for $200 million. Lattice Data applied an AI-enabled inference engine to extract dark data.

Similarly, the Chan Zuckerberg Initiative (CZI), a philanthropic organization founded by Facebook CEO Mark Zuckerberg, bought Meta for an undisclosed amount. Meta is an AI-powered research search engine startup that CZI plans to make available freely. Thus, dark data extraction technology and intellectual property developed in-house is also potentially independently quite valuable.

Data-extraction tools

There are many open source dark data extraction tools. This section shows some of the more successful tools.

- DeepDive: Stanford University developed this open source tool, commercially supported by Lattice Data. Development is no longer active with Apple's acquisition of Lattice Data in 2017.
- Snorkel: Stanford University also developed this tool. Snorkel accelerates dark data extraction by developing tools to create datasets to help train learning algorithms for dark data extraction.
- Dark Vision: This app is a technology demonstrator that uses IBM® Watson® services to extract dark data from videos, a classic example of dark data extraction.

Dark data: an untapped resource for better analytics

Dark data is the untapped data organizations routinely store during normal operations. This dark data typically remains unused because it's inaccessible to traditional relational database tools. Typically, this is because the data is in an unstructured, unusable format (for example, document scans or because poor metadata descriptions do not allow efficient searching. The quality of any data analysis depends on the body of information accessible to the analytics tools both promptly and in full detail. Dark data is, therefore, a big limiting factor.

The proportion of dark data to usable data tends to be huge. For example, in this news release, IBM estimates that 90 percent of all sensor data collected from Internet of Things devices is never used. This dark data is valuable, however, because it's data that isn't available in any other format. Therefore, organizations continue to pay the cost of collecting and storing it for compliance purposes in the hopes of exploiting it in the future.

Storing and securing dark data does have associated costs and risks, some of which exceed its value. Also, dark data can be time sensitive, and the longer the data remains inaccessible, the more value it loses. As a result, many organizations resort to human resources to manually extract and annotate the data and enter it into a relational database — an expensive, slow, and error-prone process. The advent of deep learning has made it possible to create a new breed of intelligent data extraction and mining tools that can extract structured data from dark data much faster and with greater accuracy than human beings can. The technology for these tools is independently quite valuable.

Related topics

- IBM Watson Studio
- The IBM Cloud spotlight on dark data
- The dark side of the source: Bringing unused customer data into the light
- Dark data discovery: Improve marketing insights to increase ROI