Illuminate Dark Data for Deeper Insights

Bringing more data-driven decisions to light

Tho Nguyen, Technology Director, Teradata
James Taylor, CEO, Decision Management Solutions

03.17 EB9662

TERADATA.

While organizations of all sizes across all industries are keen on becoming data-driven, most focus on only a fraction of the many types of available data. Not accessing a fuller spectrum of data, including those from "dark data"—emails, texts, images, photos, videos, and other documents—along with traditional data sources can limit an organization's ability to gain a complete picture of their customers and operations, and exclude them from game-changing insights that improve business outcomes.

This paper examines dark data and the new technologies that make it worth storing and analyzing. Dark data includes sensor and streaming data, image data, audio and video data, as well as semi-structured data (e.g., log files, survey data, notes or presentations, email correspondence, and financial statements). New analytic technologies—such as artificial intelligence, cognitive, deep learning and machine learning—are shining new light on dark data, allowing organizations to gain more business value from a wider spectrum of data.

Data-Driven Decisioning

In the race to be data-driven, most organizations focus overwhelmingly on traditional structured data that can be formatted in rows and columns, and stored in relational or columnar databases. Even with this narrow data focus, data-driven decisioning is taking over with analytics driving it across every industry.

The move toward data-driven decisioning is supported by a broad landscape of data and analytic technology—technology that is being applied to a wider range of data types beyond traditional structured and numerical data. Organizations are storing more data types, analyzing more data sources, and using these analyses more effectively to deliver better results.

Yet most organizations have some "dark" data, defined by Gartner as "…the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes…" (Gartner IT Glossary).

The good news is that new technologies and approaches are letting organizations shine a light on this dark data, putting it to work to improve data-driven decision making.

What Makes Data "Dark"

According to IDC, 90 percent of unstructured data is never analyzed. Data is considered dark when it's not being used for anything. It might be collected, processed, and stored like structured data; however, no action is taken beyond those steps to analyze, understand or use it. It is "write-only" data.

This is surprising since most companies assume that data is going to provide value at the time of its acquisition. After all, a lot of time, resources, and money is invested in collecting data, so it makes sense that it should be considered important and strategic. Data can also be considered dark if it's not gathered in a way that allows it to be analyzed and used, or simply because it's just too difficult to store.

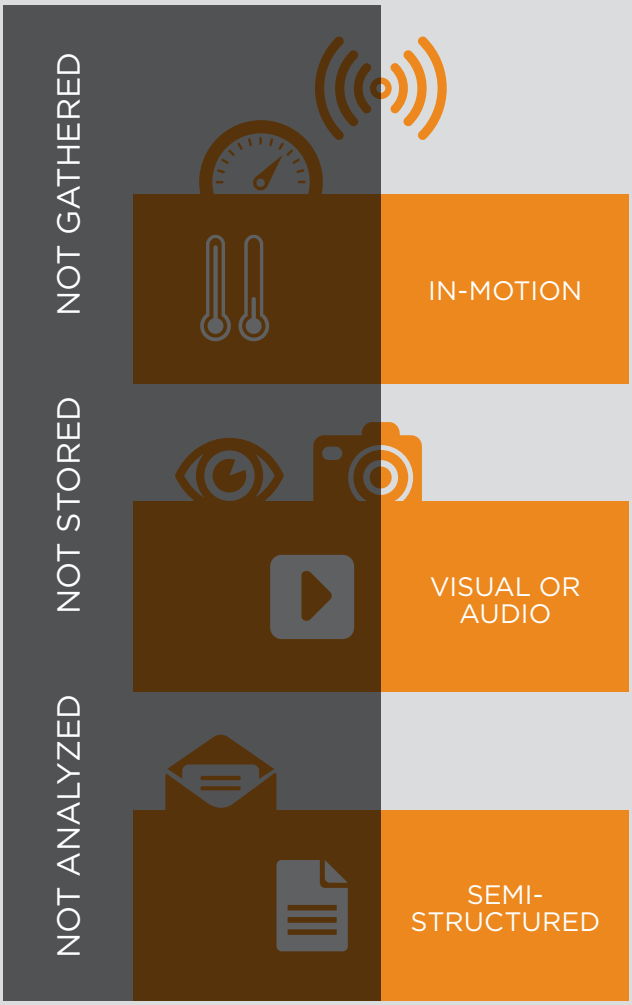
Data Not Gathered

Some data might be useful but is never gathered, such as those at an organization's end points (e.g., a customer's response to an offer, or the questions they ask when presented with product choices). The increasingly complex web of partners and subcontractors involved in many business-to-consumer (B2C) interactions can also act as a barrier to gathering data. If everyone involved in a transaction does not work for the same organization, gathering all the data about the transaction can be difficult.

Some data is not gathered because it's too complex or costly. As an example, sensors that generate data may not be connected to something that can store the data. Even when they are, conversion from real world to data storage is lossy; that is, less precise and less detailed (e.g., when sensors are being manually watched by people who then record values periodically). The continuous data available in the real world and logged by the sensor is turned into intermittent and incomplete data in storage.

Another reason some data is not gathered is simply because it's not considered to have value, and therefore not worth the effort. In this case, historically no one has cared enough about the data to consider gathering and storing it, so it can't be analyzed.

Types of Dark Data



Data Not Stored

Closely related is data that is gathered and looked at, but not stored for analysis. Even if the data can be realistically gathered, the volume or velocity of it may be a problem. High speed, high volume data may stream into an organization, but not be stored at all. At best, this data may be watched to see if a threshold is breached, or counted to see how many events there are; however, the richness of it is lost and, because it's not stored, it cannot be analyzed to see if there are deeper insights to be gained.

Data Not Analyzed

The classic form of dark data is that which is captured and stored, but never used to improve decision making. This data is the "write-only" data that organizations store out of habit, either because they hope it will be useful or think they should for compliance purposes. It can also be considered dark data if it can't be analyzed because the technology simply does not yet exist; or if it does, it's too slow to analyze the data in a useful, timely manner.

It can also be because an organization has never invested in learning how to analyze dark data; i.e., the technology might exist (and be practical) but it is unknown to the organization. Almost all semi-structured and documentary data used to fall in this category, while image, video, sensor, and other types of data still do.

Categories of Dark Data

In theory, any kind of data can be considered dark; however, it's generally the types of data that are harder to gather, store, or analyze. Semi-structured or unstructured textual data, visual or audible data, and data in motion are those types that most commonly go dark.

Semi-Structured Data

Semi-structured data is the simplest and most common form of dark data. Text data from notes, documents, or emails may be stored as binary large objects (BLOBs) or outside classic databases. Semi-structured data is often dark because it's not analyzed. Simple analytic techniques to report or graph data are of no use, and more sophisticated mathematical approaches—such as data mining, machine learning, predictive and prescriptive analytics—have generally focused on structured data. While techniques exist to analyze this data, they are less commonly used or understood.

Visual (or Audible) Data

Image, video, and sound data is often not gathered at all or, if it is, it's stored only to create a historical record. Generally, these types of data take up a lot of memory in your system. No matter how it's stored, it generally requires someone to look at it or listen to it, a time-consuming and expensive process. And even if the data is watched or listened, most organizations only analyze the summary data that the watcher or listener encodes—losing a great deal of precision in the process.

In-Motion Data

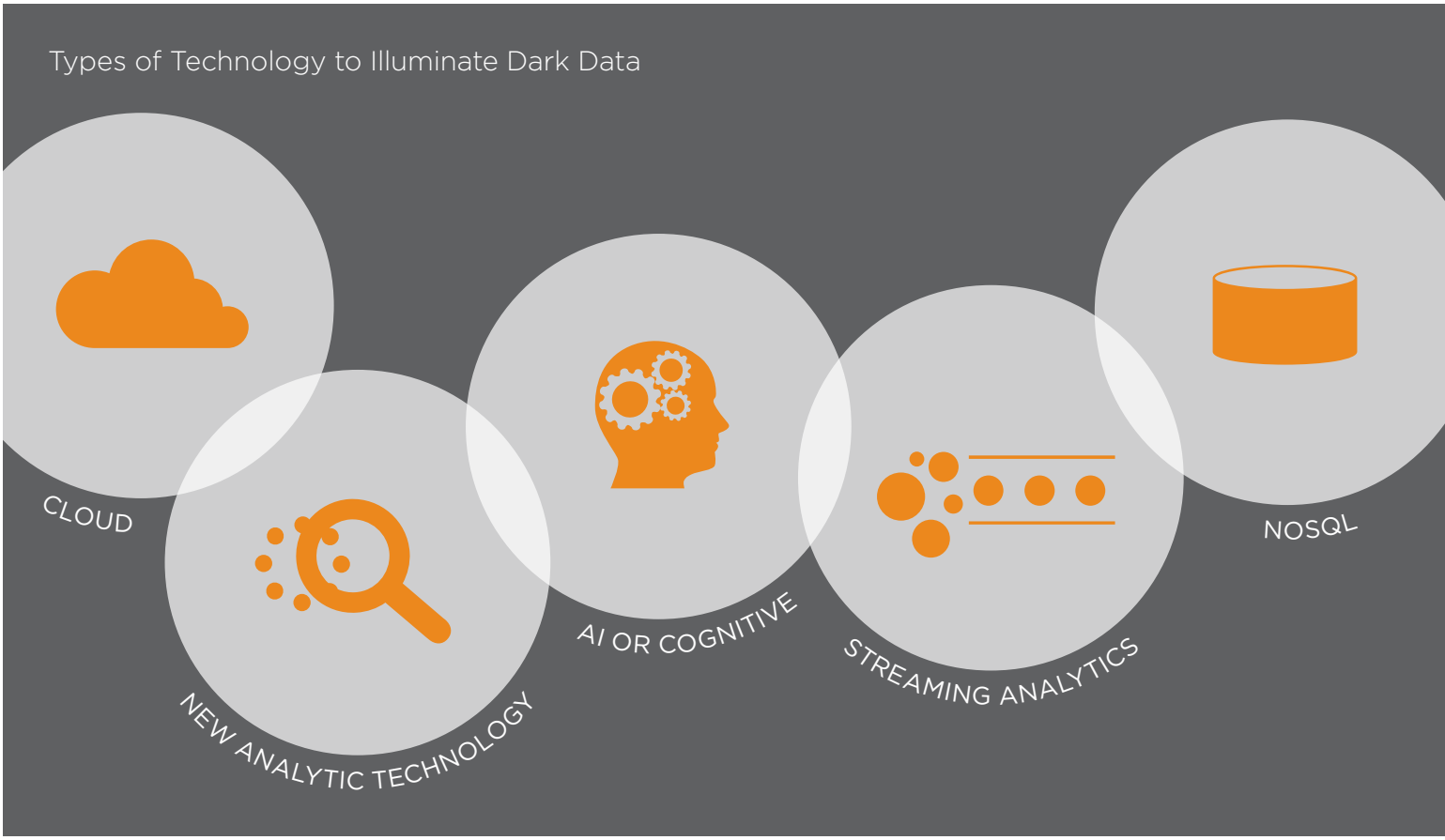
As the world becomes more instrumented and automated, there are more sensors and data feeds available to organizations. These generally produce very large amounts of data with lots of repetition, and very little variation from record to record. Imagine, for instance, a temperature sensor that transmits data every second. When the temperature is stable, 86,400 identical records would be generated every day. Sensor and other time-series data streams in, and is not generally stored for analysis. In extreme cases, only streaming analysis is possible given the pace and volume of data.

Transforming Dark Data

New technologies are making dark data increasingly accessible both directly, by enabling new kinds of analysis for instance, or indirectly by making existing approaches more scalable or cost-effective. Cloud, NoSQL, streaming, artificial intelligence/cognitive, and analytic technologies each have a role to play in making it possible to gather, store, and analyze data that would once have been dark. As these technologies become more established, best practices and services that leverage them also become more widely available, further accelerating their adoption.

Cloud Platforms

Cloud technologies allow data to be gathered and stored more effectively. Organizations that have many endpoints—locations, vehicles, devices—often suffer from delay as data from those endpoints is gathered up, transmitted, and integrated across the organization. The sheer effort involved means that many gather only summary data from those endpoints. As a result, the detailed data is often dark, available only at a specific end point and unintegrated with data from other end points.



Increasingly robust cloud technology allows data to be gathered directly from the end points; data can be stored to the cloud as it's created, ensuring that all the data is gathered up. Cloud-to-cloud integration allows data to be rapidly and cost-effectively integrated and stored. Cloud storage can give organizations access to all their data, without time-consuming and costly transmission and re-transmission of data around the organization.

NoSQL Data Storage

Much of dark data is unstructured or semi-structured data. NoSQL technologies allow this data to be flexibly and cheaply stored with "schema-on-read" approaches, allowing access to the data at a later date. NoSQL storage increasingly means that all the data an organization has can be stored so that it can be read and analyzed later.

It also means that new data sources can immediately be added to the available data without time-consuming analysis and design work. Organizations can prevent a new data source from being dark temporarily, while it's modeled and analyzed. The flexibility of applying a data schema only when the data is read allows it to be gradually illuminated as more ways to use different pieces of data become apparent.

Streaming Analytics

Some data remains dark simply because it's moving too fast to be acted on. While it may be possible to store the data for analysis later, there is simply no way to inject the results of that analysis into the data stream so that the analysis can be applied as the data flows past. In this case, the data is being analyzed but the analysis is not being used effectively.

The growing ability to embed analytics in real-time and streaming systems means that analysis can be applied even to very fast moving data that is not being stored. Streaming platforms can consume analytics and apply them to rapidly changing data, using time and event windows as well as embedded execution to effectively and analytically decide without ever having to store the data.

Artificial Intelligence (Cognitive Technology)

One of the most interesting areas for many organizations is the growth in cognitive technology. This relies not only on analysis of existing data, but on being taught. For instance, cognitive algorithms can be taught to recognize sentiment or tone in text, spot damage in images, "hear" patterns in audio, and much more.

The ability of these techniques to process natural language as well as images, audio, and video data dramatically expands the range of data that can be illuminated. Many organizations have stored these kinds of data but assumed that only human analysis was possible. This limits the degree to which patterns can be found when data volumes are high; does not allow analysis in automated transactions; and drives up the cost of analysis. With new cognitive technologies, this kind of data is available for large scale, cost-effective, and automated analysis.

> New technologies are shining light on dark data, bringing analytic transparency to your operations.

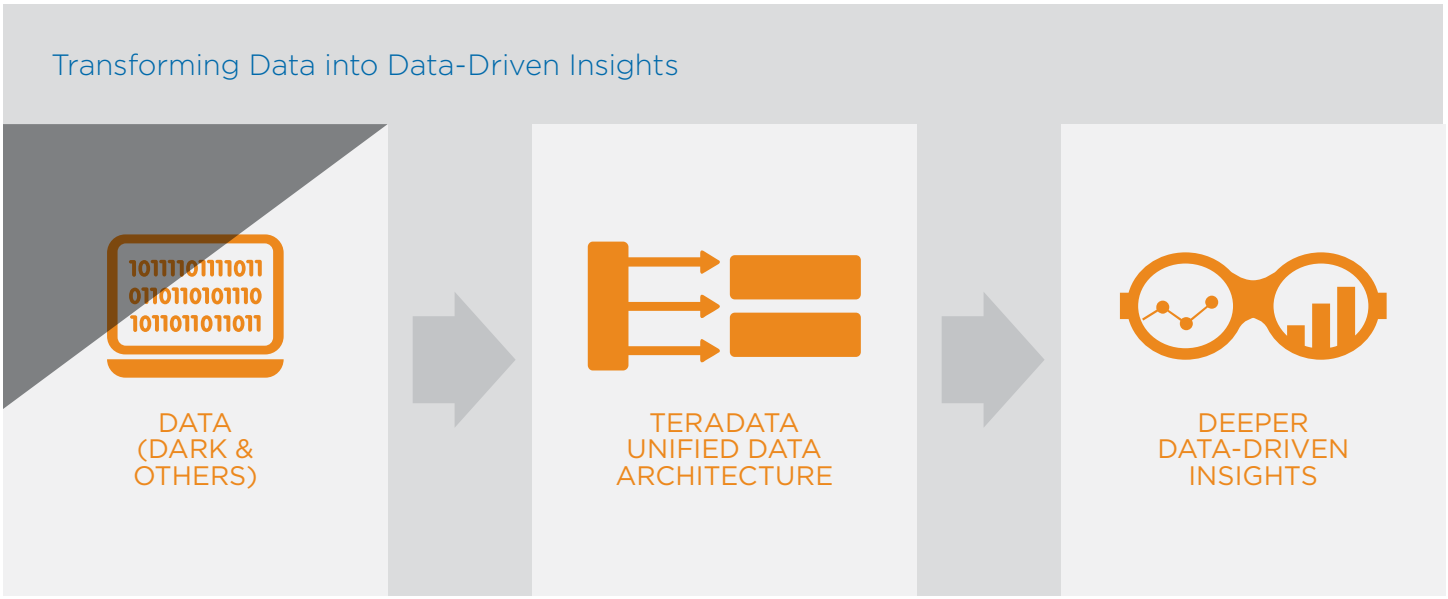
New Analytic Techniques

In addition to these specific technology improvements, new and improved analytic techniques are playing a role in bringing data out of the darkness. The ability of analytic techniques to extract structured meaning from unstructured text—using entity analysis, for instance—so that it can be combined with structured data has increased the range of text that can be analyzed. More powerful algorithms and ensemble methods that combine multiple approaches extract useful insight from larger and noisier datasets. Parallelized algorithms allow all of an organizations' data, not just a sample, to be rapidly analyzed. Volume and complexity are no longer barriers that keep data dark.

Teradata Technology

Since dark data can consume a lot of memory, it requires a scalable and high-performance platform. Within the Teradata Unified Data Architecture™ (UDA), dark data can be loaded into the integrated data warehouse, discovery platform, and/or data platform.

Teradata offers a variety of technologies and solutions that help manage and analyze dark data. For customers who are interested in the cloud option, we offer private and public clouds to complement its on-premise solutions. Once the data is stored in the appropriate platform, it can be analyzed and connected to the right business processes for deeper insights. Dark data can also be integrated with traditional structured data for comprehensive analysis. Since the data has high business value, seamless access makes it easy for business users to follow through on execution.



We've partnered with a number of vendors to complement the Teradata UDA. Teradata works with best-of-breed, innovative vendors to deliver analytics from images; integrate cognitive and artificial intelligence; and provide event stream processing to provide an end-to-end, comprehensive data management and analytic lifecycle. Teradata believes that capturing, integrating, and analyzing structured, semi-structured, and image data on a scalable, high-performance platform with the right analytics solutions delivers deeper insights to drive better business outcomes in a timely manner.

Teradata acknowledges that new technologies and techniques require new skillsets to implement and adopt. Customers who do not have in-house expertise can leverage Teradata's services to help them with strategy, assessment, and implementation. We provide complete services from data gathering and data analysis to on-going project support.

Use Cases

Retailers

The ability to analyze behavior and trends for identifying products to promote or cross-sell, predict the sizes customers will need, and more, remain critically important for retailers. But this analysis is all done on structured data. What if you could capture and analyze images of outfits your customers put together? What if you could tell that certain age groups wear certain clothes in a baggy size? Could storing and analyzing image data provide new ways to target certain age groups, and would it change what you stock? How can you leverage your brand to its fullest to reach a wider audience?

Hospitality

The hospitality industry bases much of its targeting on explicitly stated guest preferences and status. But many guests never bother to document their preferences on the hotel website, or don't have the status to get staff members to ask and record it. What if organizations could analyze social habits and other unstructured text to learn customer preferences without making them write it down? Could they customize and target the customer experience better?

Health and Life Sciences

Health and life sciences collect data from patients. Sensors in facilities and at home, combined with managed images, offer tremendous opportunities. Streaming data from sensors can be analyzed to predict patients who are heading toward a crisis in between readings, while images can be automatically scanned for potential threats so that patients can be prioritized and treated effectively.

Insurance

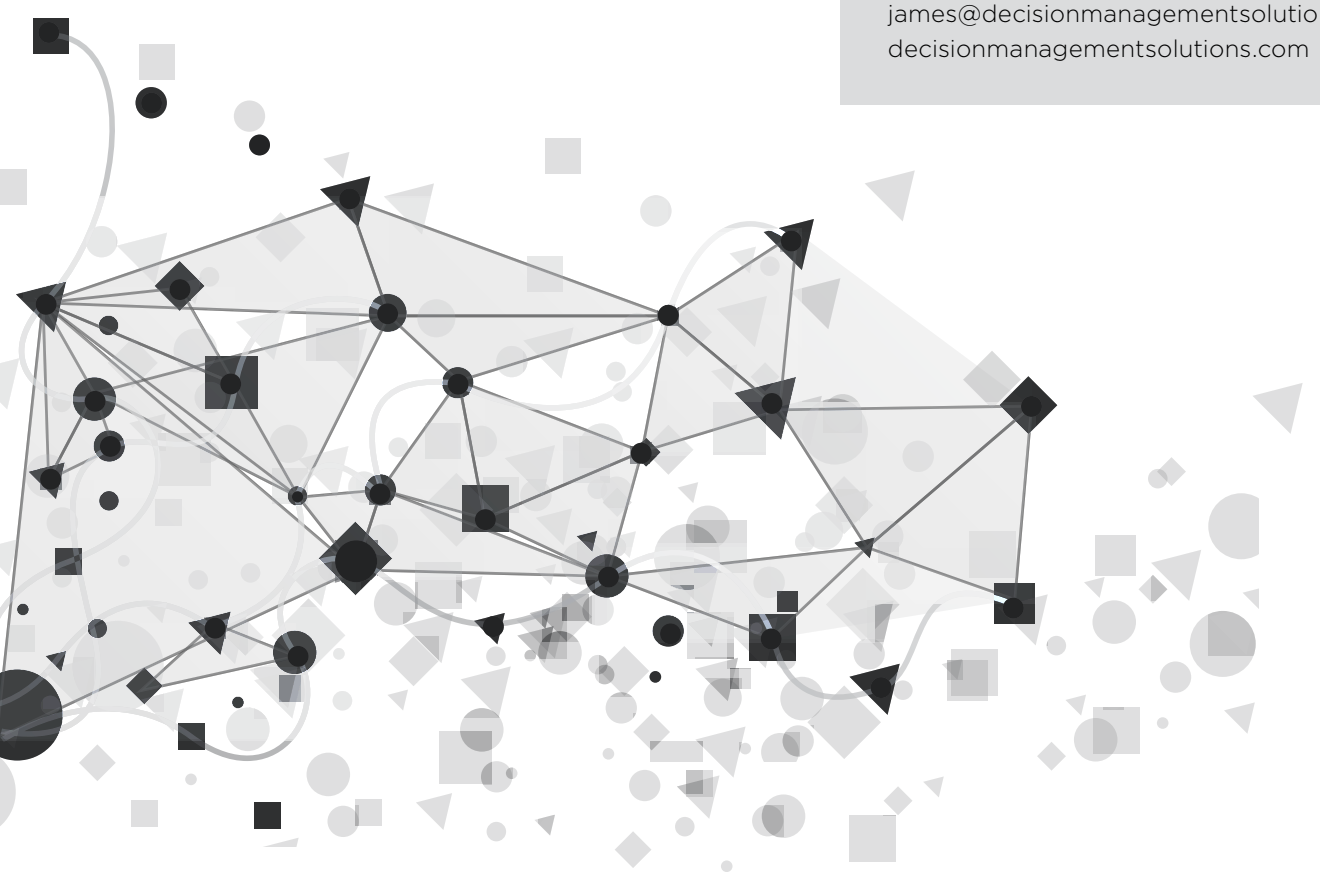
For years, insurance companies have gathered photos showing damage. New artificial intelligence and analytic technologies are able to spot holes in roofs, match damage photos to specific parts of cars, identify the age and condition of something being claimed against, and much more. With geospatial data and data from vehicle sensors, it's possible to tell who might need medical assistance, and which vehicles are safe to drive away. How might analyzing these photos change the insurance experience?

Conclusion

In today's highly competitive data-driven business environment, it's become increasingly important for organizations across all industries to capture, store, and analyze the full spectrum of available data—dark and traditional data sources—to drive better business decisions and outcomes.

When semi-structured, image, audio, and in-motion data are neglected, the potential for analytic insight is reduced, critical decision making suffers, and business value is lost. New technologies like private, public, and hybrid cloud infrastructure, noSQL data storage, streaming analytics, artificial intelligence/cognitive technologies, along with powerful new analytic techniques are illuminating dark data like never before to reveal deeper, actionable insights.

To learn more about how your organization can harness the potential of all your data sources—and optimize its business value through analytics—visit Teradata.com/contact-us.



About the Authors

Tho Nguyen, Technology Director, Teradata

With nearly two decades of experience, Tho works closely with R&D, partners, and customers to drive and deliver value-added business solutions in analytics and data management. He helps companies become more data-driven, and has recently published a book on the topic. Tho is an active blogger and speaker on data and analytics, and a faculty member of the International Institute of Analytics.

James Taylor, CEO, Decision Management Solutions

DECISION MANAGEMENT SOLUTIONS

James has been a leading expert in how to use business rules and advanced analytic technology to build decision management systems for over 15 years. He is a published author of numerous books and articles on Decision Management and decision modeling, as well as a regular speaker and blogger. James is a faculty member of the International Institute for Analytics.

james@decisionmanagementsolutions.com
decisionmanagementsolutions.com

10000 Innovation Drive, Dayton, OH 45342 **Teradata.com**

03.17 EB9662