Data Lakes and Why Business Might Want One

Building on the Teradata Experience

September 2016

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Data lakes have been a hot topic in data management circles for over five years now. While their exact architecture is a topic of debate among IT, their general shape and structure is well explained in the metaphor of a body of water where business can explore and play with a wide range of big data.

This paper focuses on the business value that can be garnered from a data lake. The approach is to examine a number of real, but anonymized, examples across different industries and to allow them to speak for themselves. Specific lessons are then generalized into a set of "Big Pictures".

The use cases are divided into three broad categories named in the spirit of the data lake metaphor: fishing for behavioral data to understand customers, swimming with customers to enhance engagement, and navigating to operation excellence to maintain engagement. These categories offer a progression in business focus from marketing, through support, to operations and production uses of the data lake.

Finally, a short, more formal appendix is provided to clarify a common question among both business and IT: what is the difference between a data lake and a data warehouse.

Contents

Sponsored by:
Teradata Corporation

www.teradata.com

Irrespective of your market, industry or function, digitalization is coming toward you with all the poise and finesse of a runaway freight train. Tomorrow's business—and, indeed, today's—is emphatically data driven. Social media pushes business action based on personal behavior and opinion. The Internet of Things (IoT) monitors and measures the entire physical world for business value. Cognitive computing will take decision making to light speed. You will need fast reactions and deep insights to survive, never mind thrive.

Allow me to offer you an oasis of calm. Welcome to Lake Data, where your business information needs are paramount. Let me share with you some stories of success in the data lake, couched in business terms, relevant to business visionaries and leaders. Use cases from diverse industries, spanning all business functions. Tales of information used to understand customers, improve processes and drive profit. The approach of this paper is to examine a number of real, but anonymized, examples across different industries and to allow them to speak for themselves.

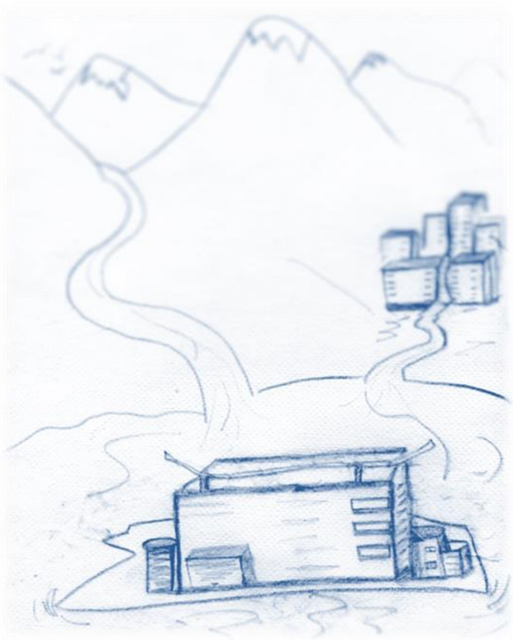


Figure 1: Metaphor for a data lake

What is this data lake? From whence does its magic arise? From the business perspective, think in terms of a metaphor. The data lake is a large reservoir of large quantities the raw data available to feed your business with every imaginable insight. It is fed by streams of detailed, crystal-clear data from the mountains of the physical world. Into it flow rivers of information, often polluted by the cities of people on their banks. In the center of this vast lake stands a manmade island: your business. There stand the operational factories that transform data and physical assets into products and the data warehouse and marts used to manage the business performance and report to shareholders and regulators.

A more formal definition of the data lake and its relationship to the data warehouse, aimed at your IT colleagues and others who may lack appreciation of metaphors (as well as poems and love songs), can be found in the appendix, "Mapping the data lake". Now, let's dive into the metaphor and recount success stories from the shores of Lake Data, based on a three-stage evolutionary model of business value:

> *As a metaphor, consider the data lake as a large reservoir of vast quantities the raw data available to feed your business with every imaginable insight.*

1. Fishing for opportunities based on ***client/prospect behavior:*** the what, when and why of clients' actions can be inferred from their growing trail of physical and social footprint data

2. Swimming with clients in proactive ***engagement:*** moment by moment, anticipating their needs, satisfying their expectations, offering deep understanding and immediate solutions

3. Navigating to ***operational excellence:*** delivering products, solutions and services of the highest quality, on-demand and tailored to ever-changing client needs

Much discussion of data lakes focuses on customers, as if sales are the only goal. However, we use the term *client* to include behavior, engagement and operations in enterprises or activities where profits are not the sole driver.

Fishing for clients in behavioral waters

Sales and marketing teams have long dreamt of perfect knowledge of their customers—past, present and future. Customer relationship management and 360° view of the customer programs were long limited by the fact that traditional systems record only direct, legally binding business transactions. Behavioral data from web logs, social media and beyond dramatically expands your knowledge of what customers were doing and even thinking before they purchased—or didn't—your product. You can discover how they use your products or services, what they like or dislike about them, and if they are considering your competitors' alternatives. Such information is priceless, provided you can store it cost-effectively and analyze it deeply in a timely manner.

In the data lake metaphor, you are fishing for insights in this behavioral data, but they are rare, elusive, and tasty only when fresh. Landing them on your island and processing them in your data warehouse may be too slow. Storing that volume of data in the warehouse may be prohibitively expensive. Dealing with it where it first arrives, in the data lake itself offers a better and faster approach to analysis and a more cost-effective long-term home for such data.

I want to look inside your head, yes I do

This line from Peter Sarstedt's famous (among those of us of a certain age) song[1] could apply to the goals of a pay-per-view (PPV) provider who wishes to optimize the delivery and viewing experience for customers, program providers, advertisers, and—of course—themselves. From targeted advertising through optimized programming to customer acquisition and retention, PPV TV offers an ideal environment to deeply understand customer wants and needs, based on observable behavior on a regular and ongoing basis.

Data on viewer behavior is available in real time from a number of sources: clickstream data from set-top boxes as well as social media likes and comments provide immediate feedback about viewers' program selection, engagement levels, channel flipping and, more importantly, the reasons for these actions. Customer care and billing records provide traditional financial and complaints data. Demographic data at individual and household levels, both in-house and third party, allow detailed individual segmentation of viewers. As such, the PPV market offers an ideal example of a data rich environment and what can be done with such a wealth of data.

Using all this information, advertisement performance can be analyzed and placement optimized to maximize PPV purchases. Highly specific advertising can be directed to individual prospects. Behavioral analysis allows programming to be adjusted to improve PPV sales and acquire new customers. Based on combined social media and viewing data, personalized recommendations can be made for upcoming programs at the most effective times during prior programs the viewer has a history of watching. Within this highly instrumented and digitalized world, this PPV provider is building a complete, 360° view of their customers.

With such a breadth and depth of data, both near real-time and historical, from internal and external sources, the first challenge is to collect it and store it. A data lake offers an obvious choice for this, as well as for the next step: consolidating and cleansing it for analysis. The cost effective processing power available with commodity hardware of the data lake also supports the in-depth analysis of data necessary in this environment.

The big picture

PPV shows how behavioral data about customers and prospects can directly drive acquisition, sales and product development. Not all industries have such a wealth of data, but the principle still applies. Be it PCs or package holidays, financial services or flowers, selling to the public demands deep insight into customer wants and needs—both overt and unconscious. This can be inferred only by observing personal behaviors before, during and after the actual sale. Behavioral data from social media, web logs and, increasingly, the IoT arrives in the data lake where it is immediately available for analysis and action. Combined with actual sales, we discover what customers like and what actually leads to sales. In the case of missed opportunities, where initial interest did not lead to a sale, the data lake holds a record of what prospects actually do and when they exit from the path to a purchase. Here, we learn what really turns prospects away.

Improved medical care through faster data use

A major medical establishment has very different needs for managing and fishing in client information, but also finds the data lake to be a good solution. Large quantities of information from radiology, operations, ECG/EKG results, clinical notes, insurance claims, etc. is collected by diverse applications and aligned to an international standard (Health Level 7). Such standardization is vital for effective data sharing and reuse across different departments. Knowing a patient's medical history early allows reduced patient preparation time, better treatment choices, and enhanced future outcomes.

However, the hospital faces a challenge: it must gather, integrate, and store this information quickly and efficiently to enable subsequent medical use, sharing and analysis. Reducing the effort and time needed to find relevant clinical and patient information frees up highly trained and experienced medical staff to spend more time with individual patients, seeing more patients, sharing expertise and training new interns—in effect, doing what this hospital is world-renowned for: curing illness rather than shuffling information.

Medical data is very different from the traditional data most businesses have used for decades. It is loosely structured and voluminous. It comes in many forms, from machine data readouts to images and documents. Only with careful data management, such as indexing, cleansing, and codifying, as well as cost-effective storage, can its value be unlocked for widespread and rapid use across the entire hospital.

Traditional relational database technologies had been used in the past to store and manage this information. However, such technology is not well-matched to this loosely structured and voluminous data. A data lake provides a more suitable and cost effective platform, providing parsing, indexing, text search, natural language processing function according to international standards for content analytics in health care.

In many cases, data lakes are promoted to reduce data storage and processing cost but, in this instance, the emphasis is on supporting doctors to better help more patients more effectively. Of course, a reduced IT cost is always welcome, but far more important is the faster and improved service to patients. Clearly defined, measured and achieved optimum visit lengths for all classes of symptoms, procedures and patients is key to best medical practice. Well-managed, indexed and contextualized information is at the heart of the business and human value to be found in this data lake.

The big picture

Sometimes it's more about knowing than selling. In the professions—from medicine to law, from management consulting to financial advice—behavioral information extends beyond customer needs to the shared actions undertaken and joint results achieved in the professional/client engagement. For clients and patients alike, this information is the foundation for better results, faster action, and broader access to expertise. Such information also becomes the intellectual property of the professional business and the basis for knowledge growth and skills enhancement. But this data is also large in volume and loose in structure. The data lake offers an appropriate model to store, manage and analyze it in detail.

Swimming with clients in proactive engagement

In the good old days, if something went wrong with a product, you brought it back to the Mom and Pop store where you bought it and the owner (who actually knew you and your trustworthiness) sorted it out. That was real service and the foundation of true client engagement and long term loyalty.

Mass consumerism broke that model: without personal knowledge and trust, sorting out problems became a game of call center cat-and-mouse. A new era in customer engagement is now dawning, with customers expecting (a lot) better service. They fully understand that today's technology—from smartphones to big data—can and should enable it. They will be unhappy with less. They demand anticipation, early detection, and resolution of potential problems.

For business, behavioral information, enables improved evaluations of trustworthiness, while the IoT allows products to report back on their usage and problems. Customer engagement can actually become proactive in this emerging environment, provided that the data and information is gathered, stored, integrated and analyzed in time. A data lake can help.

The new insurance premium

According to a 2015 report by Gartner, there will be a quarter of a billion connected vehicles on the roads by 2020, up from some 23 million in 2013[2]. As an early example of the Internet of Things, automobile insurance has been among the first industries to disrupt their long-standing business model, with the introduction of usage-based insurance schemes as far back as the mid-2000s.

A large, diversified insurer in North America has been gathering and using vehicle telematics to analyze, adjust and manage risk premiums for enterprise fleets data for more than five years. As an early adopter, their solution was largely custom-built. A modern data lake solution is enhancing and expanding the possible uses and business value delivered by this application.

Insurance has always been based on shared risk: the more fortunate subsidize their unluckier peers when disaster strikes. But insurance companies have long recognized that bad luck is only one factor. Some customers are higher risk than others because of their behaviors. Success in insurance comes from balancing exposure to riskier customers with market share. Various aggregated, statistical measures, such as address, gender, claim history and credit score have long been used to assess risk. However, real-time telematics data such as speed, acceleration and braking forces, driving times and locations from individual vehicles, combined with weather and mapping data, provide more accurate and granular predictors of risk.

Telematics also changes the insurance business models by allowing early detection of risky behavior and prediction of likely losses. Drivers can be offered individualized premiums based on monitored behavior and improved behavior can be encouraged. Mainstream drivers feel empowered to act in a way that reduces their premiums. On the other hand, cover for higher risk drivers may become unaffordable. Risk is being shared in a smaller pool and the fundamental analytical basis of the industry changes.

The data needed for this new analytics differs considerably from traditional insurance actuarial data. Telematics data arrives in large volumes, in real-time from multiple sources. It is often dirty. Its content changes as the automobile industry expands the variables measured. High-speed and flexible ingestion of this data must be followed by cleansing and integration with business transaction data. In-depth, near instantaneous analysis of vehicle usage and driving patterns along multiple dimensions delivers quantified scores and enables assessment and adjustment of insurance premiums on an ongoing basis.

A data lake offers an ideal environment for much of this detailed analytic work. In addition, it provides cost-effective storage for vast volumes of such data, allowing new models to be retrofitted to and optimized on long-term historical data.

THE BIG PICTURE

Whether its automobile insurance or home automation and security, augmented reality gaming or smart cities, real-time, geo-located sensor data is the foundation for client engagement. Behavior data, seen in the previous section, is the basis for understanding the client, but this is largely a passive exercise. Engagement allows—and may even demand—active involvement by the client. Measuring electricity consumption multiple times per hour via an IoT-connected meter allows understanding of customer behavior, but enabling him/her to actively control consumption—to save money or protect the environment—brings engagement to a new level.

SERVICE EVEN BEFORE A PROBLEM STRIKES

In today's 24x7, 365 days per year computer-driven business environment, unplanned outages of IT equipment are a major headache for all concerned. Even planned downtime is seen as impacting business results. It's hardly surprising, therefore, that suppliers of enterprise computing solutions are looking for ways to minimize or, preferably, eliminate such problems.

One supplier of enterprise servers is instrumenting its machines with sensors that continuously monitor and report on hardware and software health and performance. The data captured is sent securely from the client premises to the vendor at regular intervals day and night. In comparison to the telematics data of the previous example, this data is relatively clean: its structure and content is fixed by the vendor. However, its volumes are large and increasing, and its velocity of arrival high. A data lake offers a good technical match for these characteristics, enabling rapid ingestion and near real-time analysis.

Speed of response is vital in this environment. All large enterprise clients have service level agreements (SLA) with the vendor to ensure that their own service to customers is highly available and responsive. If a server failure occurs, there are agreed timeframes within which the problem must be resolved to avoid penalty payments being triggered. Meeting these timeframes is one of the key performance indicators (KPIs) of the service level agreement. But having the engineers notified of the problem is only the first step toward resolution. If parts have to be flown in from some distant plant, the delay can impact both customer satisfaction and the SLA. Keeping spare inventory locally helps, of course, but it also drives up costs. With early prediction of potential failure, inventory can be distributed to match availability of parts with higher risk. This lowers costs while improving both the KPI and customer satisfaction.

Another KPI is the actual rate of failures that occur among shipped machines. This is another measure that must be minimized, both for contractual reasons and simply for customer satisfaction. While engineering and manufacturing endeavor to build servers that do not fail, some incidents are inevitable. However, continuous health monitoring of servers in the field allows predictive analysis to be undertaken as the data arrives in the data lake. In essence, failure prediction models are developed based on the data patterns seen in previous server failures. Incoming data is compared to these models to detect possible indicators of upcoming failures and preemptive action initiated, getting both field staff and needed parts to the customer and avoiding costly and often painful unplanned outages.

THE BIG PICTURE

It's not only in enterprise systems that failures are becoming unacceptable. As autonomous vehicles take to the roads and skies in ever increasing numbers, equipment failures pose ever greater threats to safety. Even smartphones are becoming indispensable. Reducing TTRs and FIRs (hopefully with friendlier names!) becomes a key aspect of client engagement in a fully connected, online world. As the volumes and complexity of these devices increase, the capabilities and cost of the data lake approach become critical to stability and safety.

NAVIGATING TO EXCELLENCE IN OPERATIONS

As we've seen, the data lake enables you to build processes to monitor your prospects' and customers' behaviors and infer their wants and needs. It also offers you the ability to engage proactively with them, anticipating problems or opportunities. Knowledge is power, enabling action by you to increase their commitment. Or it may drive action by them to seek satisfaction elsewhere. So, you had better be able to fulfill their expectations!

A first thought for many businesses is how to reduce customer churn through a well-informed and proactive call center. Getting the required voluminous and diverse data from multiple sources to the call center is certainly a job for a data lake.

However, customer retention doesn't end with the call center rep. It demands operational excellence from planning and manufacturing your products (or services), all the way through the supply chain. Here, too, a data lake can provide some of the necessary infrastructure.

ANSWERING THE COMPLAINT BEFORE THE PHONE RINGS

It's commonly said that it takes three times more to acquire a customer than to retain one. For providers of mobile communications services, customer churn has long been a significant problem, given the incentives offered and the ease with which customers can switch subscription plans from one provider to a competitor. However, it's seldom a case of the incentives alone. Ongoing and unresolved customer issues often precede the decision to switch, as one major North American wireless provider has discovered… and addressed.

Most cellphone customers pursue multiple routes of interaction with their providers when a problem—such as a disputed bill—arises. They visit web sites or use smartphone apps for basic interactions such as changing plans, checking accounts or paying bills. Where problems persist, they may use online complaint forms, send e-mails or phone the call center. They may even physically visit stores to complain face-to-face with a "real person" where they originally bought their phone. All this behavior creates a breadcrumb trail of information about the customer, their behavior and actions, and their problems.

This behavioral trail is highly predictive of their propensity to take that final step and move on. Particular paths or patterns of actions over time—such as an online bill enquiry followed by two calls to the call center, and a physical visit to a store—often indicate an emerging issue that can be identified even in the early stages, especially for high risk churners. The trail can be simplified into three windows:

1. Identification Window: the provider can be proactive about the emerging issue and nip the problem in the bud before the customer is frustrated

2. Correction Window: the customer may become annoyed, but there remains an opportunity to address the issue, but quickly

3. Exit Window (or Door): it's too late; the customer is already in the process of leaving

The earlier the issue is recognized and addressed, the lower the cost to the provider (call centers are expensive) and the lower the likelihood of the customer leaving. Indeed, early reaction can even improve overall customer satisfaction and brand loyalty. The solution is to gather and consolidate sufficient data and to analyze the paths taken by the customer before leaving in a timely and appropriate manner.

The data lake provides the cost-effective storage for the large volumes of events that customers generate when trying to understand and resolve problems. With tens of millions of subscribers, even a small percentage of dissatisfied customers can drive significant volumes of data spanning multiple sources. Their interactions may span a number of months, over which detailed raw data must be retained. The data lake also offers an ideal platform on which to analyze the events over time and across channels using a wide variety of tools and methods.

THE BIG PICTURE

In any subscription-based consumer business where multiple suppliers compete, customer retention is a key process. With multiple channels of interaction, emerging customer concerns and issues can be tracked only if there is a concerted effort to gather and consolidate data from the different channels and the ability to perform ongoing path analysis of the interactions over time. The data lake provides the storage and processing to do so in a timely and cost-effective manner.

EXCELLING IN THE PHARMA SUPPLY CHAIN

Keeping hospital and pharmacy shelves stocked with life-supporting medications is a vital aspect of modern healthcare. The diversity of sources and wide distribution of receiving locations leads to a complex supply chain. It is also a highly competitive market, with several independent distribution companies operating in the space. One large distribution and services company recognized that effective data management and analysis was the beating heart of a healthy supply chain and proceeded to upgrade their information infrastructure to drive the business.

To ensure proper operation of their supply chain, the company first ensured that the data required was gathered in near real-time from all links in the process, including suppliers, distribution centers and final destinations, and even their delivery vehicles. Supply and demand can change rapidly—for example, during flu epidemics, which can spread rapidly across regions—leading to the need to quickly balance or even divert shipments, order extra supplies, and so on. Siloed data cannot support such responsiveness, so this company turned to a data lake model to create an integrated environment for all their supply chain data. Speedy onboarding and early cleansing and integration of this data were important drivers. With an improved data infrastructure, reporting and *ad hoc* querying gave easier and faster answers to supplier and customer requests and enabled internal staff to keep abreast of rapidly changing conditions.

Quick reports and reactions are seldom sufficient to keep supplies flowing and maintain competitiveness. Excellent planning and anticipation is also needed. With the data integrated in a cost effective data lake, new analytical tools and techniques could be applied directly to the data in near real-time. In developing flu epidemics, for example, weather information could be combined with geographical trends in infection rates to anticipate where specialized treatments or medicines would be in higher demand and proactively ship them to those locations.

Operational efficiency was improved and, more importantly, customer care enhanced. Speedy reactions and enriched proactivity are key in this market and provided an important competitive advantage to this company.

THE BIG PICTURE

Supply chains stretch the length and breadth of the globe in all manufacturing and supply industries and participants have been optimizing them since the early days of computing. In fact, one of the earliest and biggest data warehouse successes—at WalMart in the 1990s—sprang from their automation of the supply chain. Today, data is even more varied and voluminous, and anticipating demand and avoiding supply chain outages requires large-scale, novel solutions in addition to traditional data warehouses. Data lakes offer cost-effective storage and processing to address these needs.

NETWORKING THE WORLD WITH DATA

Beneath the world's oceans and across its continents a vast web of fiber links us all with terabit per second data feeds of everything from simple business transactions to video calls and on demand movies. On this international and intercontinental infrastructure, wholesale bandwidth providers offer a range of voice, data, connectivity and monitoring services to the familiar retailers—mobile telecoms companies and ISPs—dealing with the public and small businesses, as well as large enterprises, in countries and regions worldwide.

One such company, headquartered in Europe, offers business intelligence (BI) and analytics services to its customers (and employees) so they can have a 360° view of the performance of their signaling and data products in near real-time. This service is seen as a competitive advantage in a market where traffic and event volumes are enormous and where every minute of downtime or even poor performance may have an impact of millions of dollars on the retailers as their downstream customers voice their dissatisfaction by changing supplier.

This BI and analytics service was designed with ease-of-use and an attractive interface in mind. Most importantly, it had to be capable of handling over two billion records per day. With it, ISP and telecoms retailers monitor network performance against agreed service levels, monitor and track subscribers to identify and resolve problems in real-time, and analyze consumer preferences and trends to support targeted marketing and service expansion plans.

The solution was initially delivered on a traditional data warehouse platform. However, this market is growing rapidly. Yearly increases in data volumes are running in the order of 100-200%. New products are increasing business complexity. And fraudulent activity is on the rise. Retailers are demanding more types of data discovery and analytics over longer trending and forecasting periods. These are true big data requirements in terms of volumes and velocity, as well as growth rates. The highly cost-effective and scalable storage and processing power of a data lake was used for initial processing, BI and active data archiving while a more traditional data warehouse provided real-time KPIs and alerting.

THE BIG PICTURE

Utilities, whether in the data business or providing physical products like electricity or water, increasingly need to monitor the flows in their systems in near real-time and allow customers to understand and optimize their usage. Closing the gap between the real flows in the pipes or fibers and what customers can see and influence is vital. It completes the virtuous loop from client wants and needs, through engagement, to true partnership between supplier and consumer. Similar examples can also be found in manufacturing and even in service delivery.

This last example begins to bridge from data lake to warehouse and back. From a business perspective that may seem somewhat irrelevant. It's not. There are differences in the types of data best stored in one place or the other. The appendix gives more detail but, in summary, warehouses are more suited to data you need to be right, while lakes are for data you need to keep as you found it. More than likely, you need both.

CONCLUSIONS

Data lakes are all the rage today. They are perceived by vendors and IT departments alike as the solution to a range of size, performance and cost challenges of traditional data management solutions. Like all technology solutions, of course, they have their strengths and weaknesses. It is to the IT folks we leave the resolution of those debates in terms of logical architecture and physical implementations.

Nonetheless, data lakes in the metaphor presented here do address a significant range of business concerns and opportunities. With the ability to scale to extreme data volumes and to stretch to previously unimagined levels of performance, both at reasonable cost, data lakes allow the business to contemplate the use of more types of data, at larger volumes and faster turnarounds than ever previously possible. This, in turn, offers new and traditional businesses stimulating new possibilities for process reinvention and complete business transformation.

Data lakes offer exciting new possibilities for process reinvention and business transformation.

These possibilities fall into three categories: understanding client and prospect behaviors, proactively increasing their engagement, and driving real-time operational excellence. As seen in the real use cases described here, these opportunities span all industries and involve every business function. From marketing and sales through to daily operations, businesses worldwide can now move to new levels of data usage and information innovation.

With support from vendors and IT departments to realize the concept in physical form, the data lake provides forward-thinking businesses with both the incentive and the example to reinvent themselves in an increasingly digitalized world.

Appendix: Mapping the data lake

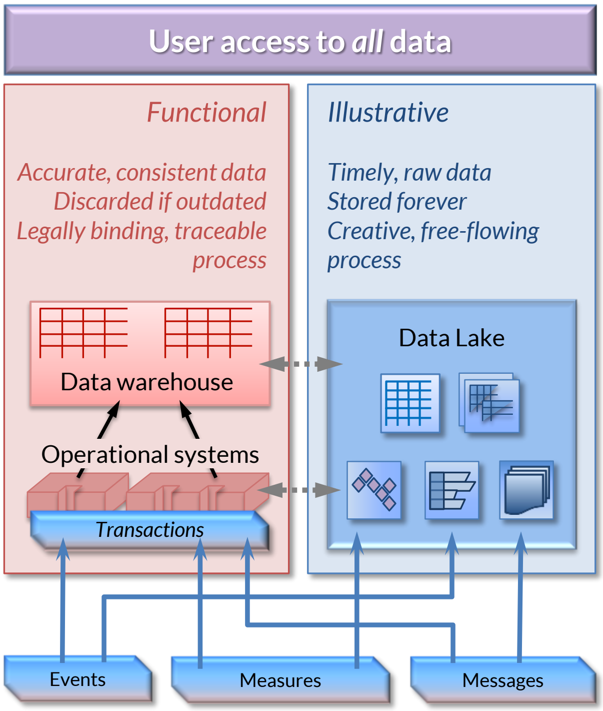
In the EMA/9sight big data 2016 survey, data lake adoption reached 67% of respondents, up from 52% a year earlier[3]. Other responses suggest that what the meaning of "data lake" varies considerably. At the most fundamental level, some consider that the data lake includes the data warehouse while others believe exactly the opposite. I believe that data lakes and data warehouses are different—but complementary—constructs.

As long-time proponents of data warehouses since the 1980s[4], both I and Teradata, who sponsored of this paper, understand the value of data to decision making and why well-managed data quality and consistency is vital to achieving that value. Such data quality allows us to distinguish between data warehouses and lakes, as well as showing why you need both constructs. In *Business unIntelligence*[5], I set out three distinct categories of data/information based on their management and usage characteristics. This approach is simplified here by recognizing that the data used in today's business has one of two purposes: functional and illustrative.

Functional is what gets business done. It begins with the collection or creation of legally binding transactions that represent real business activities like creating a customer account or accepting an order. It proceeds through the operational processes that deliver value and ends in the informational processes used to track progress and address problems. In short, it spans from Cobol programming in the 1950s to "typical" data warehouse and BI tools today. *Accuracy and consistency* of the data used is vital to functional computing; if the data is wrong, the business breaks or the regulator intervenes. Before the Internet age, these transactions were all business had to use and IT had to manage.

As web commerce, social media, and the IoT grew, it became clear that there is "rawer" data from which transactions arise. This data/information—now all digitized and potentially collected—consists of events (e.g. a click on a website), measures (the speed of your car) and messages (everything from Tweets to videos). Such data and the processes using it are *illustrative*. They allow inferences about what is happening in the "real world" beyond the business, and are the basis for predictive analytics. Data *timeliness and rawness* is central to illustrative computing; delays or summarization degrade the value of the analytics possible.

Figure 2: Locating the data lake



As shown in figure 2, these functional and illustrative purposes, with their opposing data characteristics and uses lead to a meaningful conceptual architecture that defines the shores of the data lake. Raw data—in the form of events, measures and messages—are ingested into the IT systems of the business. Vast quantities of raw data may be stored in the data lake as the basis for analytics. Teradata's definition of a data lake as *"a collection of long term data containers that* *capture, refine, and explore any form of raw data at scale, enabled by low cost technologies, from which multiple downstream facilities may draw upon"* aligns perfectly with this picture.

Data that forms the basis of legally binding business procedures is crafted into transactions by traditional operational systems and made available for decision making through the data warehouse. This separation of concerns keeps data and processes that must be well managed for business continuity and legality apart from those that require less management but allow more creativity. A data lake supports these latter needs, a warehouse the former. For business users, this separation of storage is hidden and managed by virtualization tools such as Teradata QueryGrid and metadata-based approaches. Deep links (the dotted arrows) exist between the two environments for specific business needs such as prescriptive analytic approaches that are becoming more prevalent.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His most recent book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *was published in October 2013.*

Barry has over 30 years of experience in the IT industry as a consultant, manager and distinguished engineer. As founder and principal of 9sight Consulting in 2008, Barry provides strategic consulting and thought-leadership to buyers and vendors of BI and Big Data solutions. He is an associate editor of TDWI's Journal of Business Intelligence, and a regular keynote speaker, teacher and writer on all aspects of information creation and use.

Image credits:
p1: Ingrid Hoffman; p3: mihtiander/123RF; p5: lightpoet/123RF; p7: dechevm/123RF

[1] Sarstedt, Peter, *"Where Do You Go To (My Lovely)?"*, 1969

[2] McCarthy, Niall, "Connected Cars By The Numbers", (January 2015), http://www.forbes.com/sites/niallmccarthy/2015/01/27/connected-cars-by-the-numbers-infographic/#3e46893b29ce

[3] Results from this survey will be available at www.9sight.com/resources from October 2016.

[4] Devlin, B. A. and Murphy, P. T., *"An architecture for a business and information system"*, IBM Systems Journal, Volume 27, No. 1, Page 60 (1988) http://bit.ly/EBIS1988

[5] Devlin, B., *"Business unIntelligence—Insight and Innovation beyond Analytics and Big Data"*, (2013), Technics Publications LLC, NJ, http://bit.ly/BunI_Book