# Teradata Data Lab

Stephen Swoyer

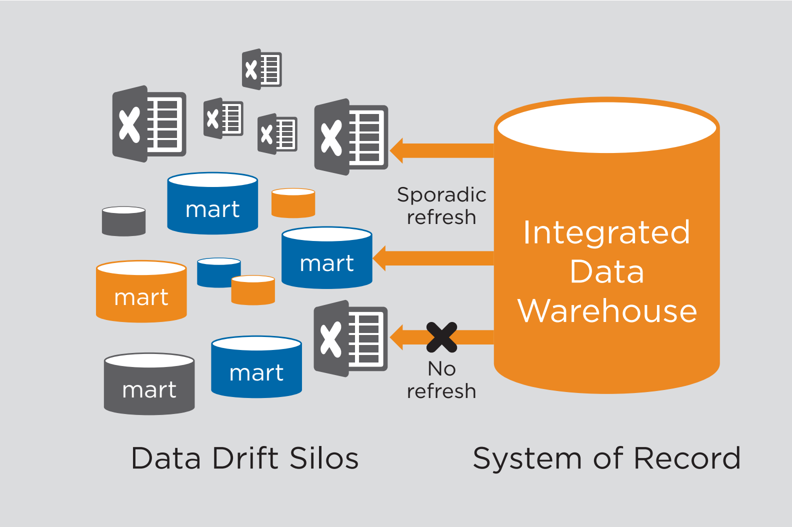
03.16 EB9338 DATA ANALYTIC ENVIRONMENT

## Table of Contents

Many organizations build analytic environments using a template that was first codified in the early-1990s, despite the availability of other, arguably better ways.

The analytic environment is usually a separate environment: separate server for processing, separate storage for data, copies of data maintained locally, separate user access and tools. The environment might be a standalone analytic data mart, or it might be a sandbox—i.e., an isolated area within the warehouse environment that contains the data which is needed by analysts.

Standalone data marts and sandboxes are used because they give business departments control. The budget is contained within a department or line of business and analysts have their own dedicated resources. Analysts control what data they keep copies of; their freedom to act—their agility—isn't constrained by the traditional IT bottleneck of provisioning data or resources. The tradeoff these environments make is one of local control versus increased complexity and inefficiency in the broader organization. The complexity is a function of locally maintained, disjointed copies of data. When analysts add data that is not delivered through common infrastructure, they create data silos. Over time, each silo drifts further from its neighbors as changes and data accrue. Over time, too, standalone sandboxes proliferate, with each new instance introducing more servers, storage, and software to manage. This adds to the complexity of managing and maintaining the operational environment and becomes more and more of a drain on the business department's budget. These rogue marts and sandboxes also tend to be slower than working with IT because their hardware, software and data must be provisioned by analysts or the department's staff. Once data is loaded, maintaining that data becomes an operational burden for which few analysts are prepared. Because of the initial gains in agility, however, analysts typically deem these tradeoffs to be acceptable. Over time, increasing complexity and costs call these tradeoffs into question. From the perspective of the business, however, IT's answer—"use a data warehouse"—is rarely palatable.



Technology advancements provide new answers to these age-old problems of local versus central control, costs, and efficiency. Teradata® Data Lab is one such answer. It comprises an agile, do-it-yourself environment in which analysts are free to provision their own data and resources. It is possible to make a central data resource appear to be multiple, independent systems. In contrast to the analyst-controlled standalone mart, Data Lab provides a centralized, virtualized environment for analytics projects.

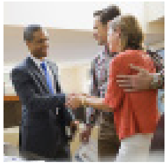
The concept is to give departments control over their own resources, carved from the larger platform. Not only resources, but also provisioning of production data that is already housed within the data warehouse. Within a data lab, analysts can provision space, resources, and data they need from the central environment. They can also load their own data into the data lab.

This solution allows the data warehouse team to provide resources, manage data security and access, and administer systems and software—tasks they are good at—without imposing their data models or views on the "right" way to achieve analytic goals. The answer for analytic environments isn't to put all of the control in IT, as with a single data warehouse for all purposes, nor is it to put all the control in the business, but to divide the responsibilities in a sensible fashion.

Data Lab is a pragmatic solution for dividing responsibility across the organization, with IT and the business departments each playing a role. It supports an agile analytic experience for analysts, data scientists, and other self-service end users. Users can perform the types of analyses they want, on the data they want, when they want, without being obstructed by IT. From IT's perspective, the Data Lab environment has the capacity to be governed, secured, managed, and scaled.

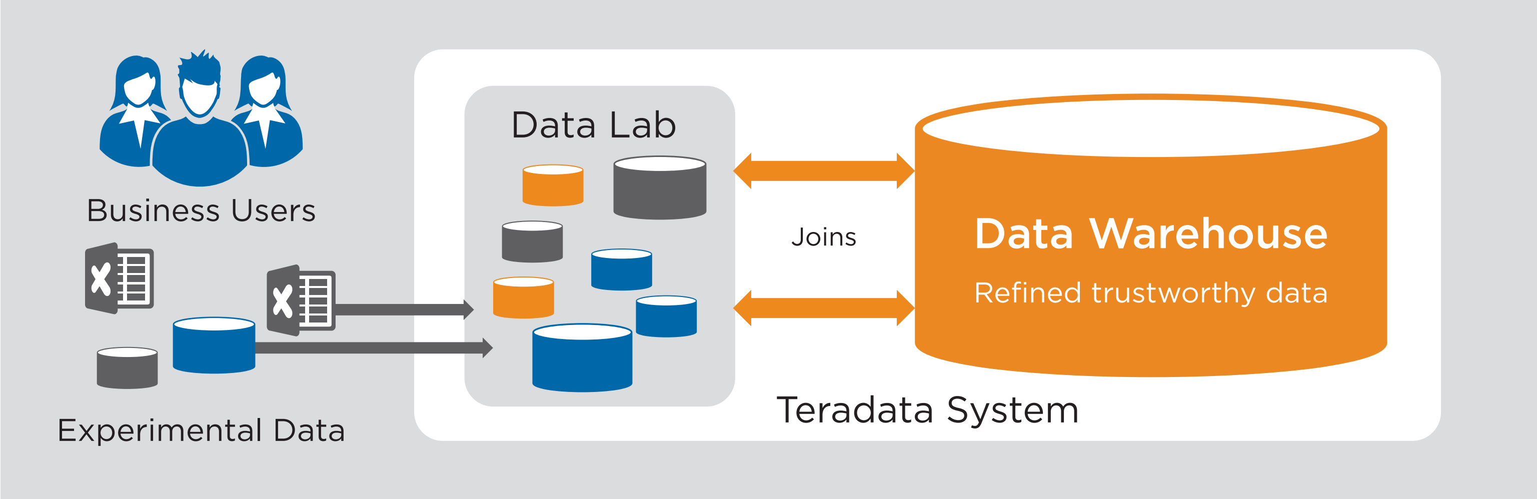
## Case Studies

### North American Financial Institution



A prominent North American financial institution made Teradata Data Lab the underpinnings of its next-generation Agile Analytics Facility (AAF) project.

This financial institution, a member of the Fortune Global 500, intended that Data Lab would function as an agile discovery environment for the research, discovery, testing, and promotion to of analytical insights. It expected to realize agility benefits in at least two respects: **first**, it



anticipated that AAF's Data Lab-based underpinnings would simplify and accelerate the process of creating analytic data sets which aid in the development of predictive models, and other critical prerequisites for analytical development. **Second**, it expected that a mature AAF, anchored by Data Lab, would promote a significant degree of agility for business users. For example, Data Lab permits business analysts to self-provision data from the production Teradata Data Warehouse and external data—to the degree that analysts are both skilled enough and willing to do so—to prepare it for analysis.

The financial institution determined that Data Lab could provide a *governed* self-service substrate for its business intelligence (BI) driven data exploration and discovery practice. Used by itself, the BI tool offers a comparatively limited data management and governance feature set. Used in conjunction with Teradata Data Lab, the financial institution could optionally enforce the same security and access-control restrictions in Data Lab that exists in the parent Teradata Data Warehouse environment. In this regard, business and IT administrators can configure access to data (i.e., database records) on a per-user or per-group basis. In turn, business analysts have access to this data—and this data alone. Data isn't physically moved out of the data warehouse environment; it is made available, instead, in the Data Lab environment, which means analysts are able to work in real-time against production data, as well as benefit from the database's massively parallel processing (MPP) data processing engine. Because Data Lab "lives" in the Teradata system, IT can enforce access control restrictions, record-level access restrictions, auditing, traceability, data lineage tracking, and other critical governance mechanisms—or can delegate this responsibility to the users, teams, groups, or units that own a particular instance of a Data Lab.

In addition to realizing agility benefits, the financial institution expected that the implementation of the AAF and Teradata Data Lab would result in significant cost savings.

Prior to implementing Data Lab, for example, the financial institution had maintained a physically separate analytic sandbox for business analysts. This sandbox environment was physically separate from its Teradata Data Warehouse®. It was likewise separate from the financial institution's third-party BI reporting and analytic infrastructure. Business analysts duplicated this infrastructure—along with third-party data profiling, data cleansing, and data preparation tools—in the sandbox environment. Because the financial institution's sandbox environment was physically separate from its data warehouse, business analysts typically worked with IT personnel to develop and implement the ETL jobs that populated their working data sets. These ETL jobs typically ran on an overnight basis. In the aggregate, they involved the nightly extraction of several million rows of data out of the data warehouse. Each night, this data was extracted and loaded into the BI sandbox server, at which point it was prepared for use in analysis.

The financial institution's analytic sandbox imposed **first-order costs** in the form of **(1)** additional BI software licenses; **(2)** redundant hardware, both with respect to the sandbox itself and to the extra storage and compute capacity required to accommodate users and their redundant, overlapping, unmanaged workloads and data sets. First-order costs came, lastly, via **(3)** the IT development and support expertise required to manage and maintain the sandbox environment.

The BI sandbox imposed **second-order costs** in the form of the labor-intensive, non-self-serviceable, unmanaged processing by means of which data sources were provisioned, allocated, and maintained. Just as important, its workload management feature set was comparatively impoverished: concurrency was an issue, such that the sandbox had to be over-provisioned to support concurrent users. In a typical Teradata Data Warehouse environment, by contrast, customers make use of Teradata's Active System Management™ (TASM) facility to manage user concurrency and to optimize workload processing. Finally, there was no automated means by which to monitor, manage, and *de-provision* jobs, resources, and user accounts. In too many cases, unnecessary ETL jobs weren't purged, even if no longer required by the analysts who first commissioned them. (For example, when an analyst left, changed positions, or finished with her project.) In the same way, a user's working data wasn't always purged when he or she left, either. Furthermore, models built in the SAS® sandbox had to be redeveloped to run in the Teradata Database® once proven accurate.

Implementing Teradata Data Lab permitted the financial institution to eliminate costly analytic sandboxes, along with the redundant BI software used in conjunction with that environment. This financial institution also used Teradata Data Lab to revamp its software development lifecycle (SLDC) for reporting and analytics.

In both cases, the adoption of Data Lab was consistent with improvements in agility, time to market, and time to value. Teradata Data Lab exposed self-service features to business analysts and enabled IT personnel to implement a more flexible test and development regimen. For example, the use of Data Lab helped accelerate the process of data discovery and data preparation by 85 percent. The Data Lab model also permits business analysts to iterate much more rapidly: the analytic development lifecycle was accelerated by 66 percent, according to the financial institution.

"The big goal from the business side was to improve their time to market. They wanted to be able to accelerate their insights," said Lesley Scholes, a Teradata liaison who helped assist the financial institution with its implementation of Data Lab. "To be able to run more analytics in more of a self-service manner was an important criterion for them, [along with] the lower cost of ownership because they wanted to get rid of the duplicate environment. They wanted to have a feeling of increased agility and self-service, they wanted to feel as if it was something they didn't have to go to IT for."

## Online Retailer

A prominent online retailer uses Teradata Data Lab to eliminate a separate sandbox environment and to promote an agile environment for data scientists, business analysts, and other skilled users. Previously, the retailer had maintained a single large sandbox for analysts and other self-service users. Database administrators termed this the "Working-database".

The Working-database was a source of ongoing tension between IT and the line of business. Teradata database administrators (DBAs) created the Working-database as a *laissez faire* environment in which self-service users could work more freely than in the enterprise data warehouse. In the governed, more tightly controlled enterprise data warehouse environment, users felt that IT imposed too many restrictions on their ability to access, integrate, and experiment with data. At the time, Teradata itself did not offer a feature such as Data Lab. Moreover, the retailer's IT organization lacked the resources to build, deploy, and maintain a functional, agile, withal manageable alternative.

In practice, the Working-database was effectively ungoverned. DBAs did not have the ability to impose meaningful limitations on user storage, nor to set expiration limits on user or project accounts. Users were likewise free to import data from the Teradata Data Warehouse at will, with little effective security or governance. The result was "catastrophic" says a DBA who led the retailer's implementation of Teradata's Data Lab. Over time, users created *thousands* of tables in the sandbox database, some of them just a few gigabytes in size, others larger than 1 TB. The analysts who used the Working-database were constantly clashing with the DBAs for additional space. "The space was never enough. There was always an out-of-space battle with the DBAs. If they ran out of space and couldn't get more, analysts used to just drop each other's tables," the database administrator said. In practice, the fact that one analyst would eliminate another analyst's tables could even go unnoticed. The Working-database was also effectively *unmanaged*—i.e., not monitored; not audited; infrequently, and never systematically, purged—which meant that most of the tables it contained were orphaned. The upshot was that an analyst could plausibly conclude that the tables she planned to eliminate weren't being used and weren't otherwise of importance. "These tables would be created by analysts who would then leave the organization. And out of these thousands of tables, only a few were important," the database administrator said.

Nor did the retailer have a governed, reusable, self-serviceable means to make production data from its data warehouse available to users in the sandbox environment. Instead, data had to be extracted and loaded into the sandbox environment, typically as part of a nightly batch routine. Not only did this lead to bloat and redundancy, it was, moreover, fundamentally ungovernable. "We didn't have any access control. What we had was an organization-wide database, with no masking of customer information," the administrator explained.

The use of Data Lab also permitted this retailer to simplify its ETL development and to accelerate the rate at which it could provision new data sources for information consumers on an enterprise-wide basis. The retailer's experience suggests that Data Lab can promote agility for data scientists and business analysts. It likewise illustrates the benefits—such as faster responsiveness, greater flexibility, and more congenial IT/line-of-business relations—that typically accrues to an IT organization by virtue of using Data Lab.

In practice, the retailer found that Teradata Data Lab permitted it to realize a highly iterative analytic development process. The Data Lab model makes it easier for data scientists and business analysts to reliably integrate—i.e., transform, engineer, or wrangle—data from the Teradata Data Warehouse with data from other sources, including social media. It likewise permits data architects, data modelers, data scientists, and other skilled users to more quickly build, test, and refine their data models and predictive models. Analytical development is to a critical degree a function of *rapid* iteration: i.e., of the rate at which analytical insights can be researched, discovered, tested, and refined. The more quickly data scientists and business analysts iterate, the more rapidly analytical insights are perfected and productionized in the larger information enterprise. Finally, the retailer found Data Lab a useful context in which to prototype, test, refine, and—ultimately—productionize data models, ETL jobs, and other IT artifacts. In this regard, the Data Lab model promotes flexibility and self-service for IT practitioners, too.

An added bonus was that Teradata Data Lab made it possible for the retailer to accelerate the rate at which it provisioned new data sources. Analysts and the lines of business were frustrated with the amount of time it took for IT to deliver new data to the users. At the beginning of the analytical process, data first had to be provisioned—typically by data modelers and ETL programmers, who worked in tandem with analysts to produce new data structures. But the retailer used a waterfall-based development scheme, which is a time-consuming, front-loaded process. By shifting its ETL development into Data Lab, the retailer was able to rapidly prototype, test, and refine ETL data flows. "The way we were doing it was mostly based on waterfall methods, which meant that if you needed to get any data inside the data warehouse, you needed to get the requirement analysis done, get the data modeling architecture done, and then, finally, the ETL would be built," the administrator said.

"In Data Lab, you're able to develop against production data in the Teradata Database, which accelerates this process. It's also better for the business users. In the old approach, they would not be able to see data until two to three weeks after the ETL was finally built. Afterward, they would often come back and say, 'This data doesn't look right on test.'"

## Research and Risk Corporation

A prominent purveyor of business research and risk management services is using Teradata Data Lab to promote self-service use cases for business users, as well as to improve its ability to govern and secure the Teradata environment. This company had previously maintained as many as eight different analytic sandbox environments for different internal constituencies. The most important of these sandboxes functioned as a single large, ungoverned, effectively unmanaged sandbox for all internal users. These sandboxes were unplanned. They were created independently and at various times to address the needs of users. The sandboxes were difficult to secure and govern because there were comparatively few controls built into them. A single user or department could change any of the data in the sandbox. "We needed to get the sandboxes under some kind of control," said a database administration lead with this company. "We wanted to build some kind of functionality into the system management rule set so that these sandboxes couldn't take over the system with a bad product join, for example."

The primary issues with the use of the original sandboxes were security, governance, and performance. The IT organization wanted to wean its user constituencies off of the unmanaged sandbox environments and shift them into the Teradata Data Lab environment. In the context of Data Lab, database administrators could facilitate access to production data in the Teradata Database, as well as impose limitations on sandbox storage and compute capacity. These limitations could be adjusted up or adjusted back down in response to user needs. In the same way, IT could manage performance via Teradata's Active System Management™ (TASM) facility to ensure that users working in Data Lab didn't compromise or otherwise impact the performance of the Teradata Data Warehouse.

The organization has shifted several users and groups to the Data Lab environment, although it has not yet achieved its goal of retiring the unmanaged sandboxes. The organization has likewise discovered a new or unanticipated use for Data Lab, as an ETL development, prototyping, and testing environment. The first such project is for a customer support application, with additional domain-specific ETL applications/projects planned.

The organization decided to implement Teradata Data Lab because it would permit users to manipulate and test sandbox data in combination with production data—but in a secure, isolated environment. The organization likewise expects to derive administrative and performance benefits in deploying Data Lab. In comparison to Data Lab, which runs in the context of the massively parallel processing Teradata Database itself, their traditional sandboxes are slow, resource-intensive, ungoverned, and (by virtue of the process by which data is extracted from the warehouse and loaded into the sandbox) dated. The organization also expects that Data Lab will foster a collaborative discovery experience as users explore data in the largest of the sandboxes, identify critical or potentially valuable information that hasn't yet been promoted to the data warehouse, and work with IT to bring it in.

## Analysis

Teradata positions Teradata Data Lab as a pragmatic option for promoting a governed and managed self-service experience for data scientists and business analysts. It says Data Lab helps IT organizations to achieve a pragmatic balance between the twin priorities of agility and governance. Agility and governance aren't mutually exclusive; nor are they irreconcilable. In practice, they're inversely related, such that any attempt to promote agility usually entails the relaxation of one or more mechanisms of governance, and vice versa.

As any exasperated-analyst will attest, a top-down mandate to promote governance invariably restricts her capacity to do what she wants when she wants to do it.

### The Challenge

The challenge for IT organizations is to promote as much agility as is possible in the context of a disciplined and governed data management regimen. This means supporting and enabling the activities of different constituencies of self-service users, foremost among them data scientists and business analysts, to the greatest degree possible.

From the perspective of the self-service user, the enterprise data warehouse is an insufficiently agile environment. This isn't because data warehouse architecture is in some fundamental way incompatible with self-service. Nor is it because data warehouse architecture isn't compatible with the kind of agility championed by self-service end users. It's because the data integration and governance processes that manage the data warehouse are incompatible with the laissez faire experience most self-service users would prefer.

Several normal data warehouse processing tasks constrain the *laissez faire* experience. First, being required to fully transform and cleanse data before loading it into the data warehouse means that experimental data is all but excluded. Hence, the analyst cannot do her job. Security controls often slow her down should she want to upload experimental data. The administrators first have to define tables to hold that data and grant security permissions. This hampers spontaneous data exploring. Last, she cannot reformat production data at will to fit predictive analytic algorithms. All of these features and services combine to constrain the user's freedom and (depending on the priority accorded to the self-service user's workloads) also limit the rate at which her analyses can be processed.

This, then, was the disconnect that led to the proliferation of spreadmarts—what industry analyst Wayne Eckerson famously dubbed "spreadmart hell." It was as a consequence of this disconnect that the agile movement in BI and analytics emerged. It promised users more power, more freedom, and—most important—a kind of self-determining agency. Spreadmarts in the form of shared Microsoft Excel® spreadsheets, Microsoft Access® databases, or rogue Microsoft SQL Server® databases are nothing new. In a sense, standalone data discovery and data visualization tools are just the latest incarnation of this phenomenon. This is not to diminish the usefulness and impact of these technologies: visual data discovery is a valuable tool for data analysis. There's no disputing that. In practice, however, the generalized use of visual discovery tools tends to have a spreadmart-like effect. Data discovery tools, like spreadmarts, are relatively easy to acquire, deploy, and use. Data discovery tools, like spreadmarts, are labor-intensive to manage and maintain. In practice, the use of data discovery tools also results in spreadmart-like siloes.

In isolation one-off or stop-gap spreadmarts are a necessary evil. They enable business users to accomplish analytic tasks that IT can't or won't deliver quickly enough. In the short term, they permit a *laissez faire* agile user experience. However, spreadmarts eventually limit agility because.... For this reason, spreadmart-like technologies permit what could be called a *pseudo-agile* experience. They're "agile" in a context, a bubble, that ignores, which is blithely indifferent to, the constraints of the real world.

Heavy-handed governance is the primary cause of business-user frustration or dissatisfaction with the data warehouse. Again, this isn't something that's endemic to data warehouse architecture itself; it's rather a function of an overly restrictive governance regimen. Changes to the data warehouse can be implemented in seconds or minutes—*provided* DBAs aren't constrained by governance policies or processes. In too many organizations, BI Competency Center (BICC) governors have become *de facto* data jailers. The upshot is that changes to the data warehouse can take weeks or even months, because the capacity of the organization to respond to changing conditions and circumstances is subordinated to governance processes. This isn't business agility. This is obstinacy.

## Agility, Properly Considered

Nowadays, the do-it-yourself analytical environment can consist of anything from a consumer-off-the-shelf (COTS) relational database—Microsoft's Access and SQL Server databases were and are popular options—to one of several OSS databases, to a desktop-based data discovery tool. From the self-service user's perspective, any of these offerings can comprise a highly workable solution. Pair a dedicated (R)DBMS with Microsoft Excel or PowerBI—or, for that matter, any of a slew of available data discovery tools—and you've got a usable do-it-yourself analytic sandbox.

And that's the problem. Independent analytic sandboxes place the onus for data preparation on the self-service user. Some classes of users—data scientists, for example—might prefer to prepare their own data sets; for business analysts and other self-service users, however, data prep is usually a tedious and counter-productive exercise. In addition, the generic sandbox environment does not provide a mechanism for standardizing data extracts (data flows) as ETL jobs, or for otherwise promoting the reuse and management of data sets and data extracts. It is error-prone. Finally, this approach gives priority to the individual analyst, working in isolation. It doesn't recognize, valorize, or promote the sharing and collaboration of data, research, and results.

A realistic way of thinking about agility is as the capacity to operate—to *act*, to *do*, to *effect*—in the context of limits and constraints. In a sense, too, what we mean by "governance" is what acts against, impedes, controls, limits, etc, the agility of the self-service user. It's pointless to talk about agility without framing this discussion in the context of a set of constraints or restrictions that delimit what it means to be "agile." For this reason, it is dishonest and pernicious to conceive of agility in an abstract, unqualified sense, as so many self-service BI vendors do. The "agile" user experience these BI and analytic tools permit is a function of their paying short shrift to the restrictions and limitations of the real world. These tools are "agile" in a fantasy world that does not align with the business priorities, policies, and regulatory requirements. On this same basis, the classic sandbox is a no less "agile" environment, as is the venerable spreadmart. All of these tools are agile in the sense that they permit the business user to circumvent reasonable and pragmatic restrictions—or to operate without any knowledge or awareness of said restrictions. This isn't agility; this is willful denial. It is abnegation of responsibility.

It is, to put it differently, an eyes-wide-shut approach to self-service. One popular caricature of the data warehouse was of a system so encumbered by restrictive policies and policy-enforcement mechanisms as to preclude its effective use by the very people for whom it was designed, funded, and built. The eyes-wide-shut approach to self-service is no less of a caricature. It's predicated on the belief that "agility" connotes the utter absence of restriction, control, or limit.

The *naive* self-service analyst would prefer absolute agility; the naive IT organization would prefer absolute governance and manageability. The pragmatic incarnations of both recognize that viable, sustainable, and agile user productivity must encourage the fail-fast development, testing, and refinement of hypotheses. It must promote exchange and collaborative research among self-service analysts. It must make it as easy as possible to reuse and productionize analytical discoveries. It must, lastly, have the capacity to be governed and managed to pacify auditors.

## Teradata Data Lab as an Alternative to Naive Self-Service

Analysts do better work and are able to produce better business outcomes when their own preferences for agility are balanced, pragmatically, with the sensible needs of governance and data management.

Interviews and briefings with almost a dozen customer references demonstrate that Teradata Data Lab can support an agile user experience that permits self-service users to work faster and to be more productive. It achieves this in several ways. **First**, Data Lab "lives" inside the Teradata system itself. This eliminates many spreadmarts and ensures the use of high-quality production data. It also makes it easier for analysts to combine experimental data with production data warehouse data. Customers cited this as one of Data Lab's most valuable features. **Second**, it extends the capabilities of self-service data analysis and data preparation tools: customers describe the Data Lab analytic environment as an agile, self-serviceable complement to their existing self-service front-end tools. Customers say the Data Lab model helps standardize and simplify the process of getting new data and applications into production. Users are able to experiment in the Data Lab environment with proof-of-concept projects that, once finished, can be promoted into production. **Finally**, the Data Lab environment can be governed and managed on an extremely granular basis. Implicit in this is the capacity to *relax* many governance requirements for certain users, groups, projects, and so on. Customers say that Data Lab permits them to strike an extremely fine balance between the needs of self-service data scientists and business analysts and the priorities of governance. Simply put: if you're a large Teradata shop, you can and should be using Data Lab.

### In-Database Processing Accelerates the Analytic Lifecycle

Data Lab enabled the financial institution, a large North American financial institution, to accelerate the process of data discovery and data preparation by as much as 85 percent. The company made Data Lab the foundation of its next-generation Agile Analytics Facility (AAF) project, an initiative to develop a sustainable self-service analytic practice for data scientists, analysts, and other savvy users. In practice, the Data Lab-powered AAF simplified and accelerated the process of developing data models, predictive models, ETL data flows, and other critical analytical artifacts.

The financial institution's case is a good illustration of the different ways in which an analytic environment such as Data Lab can extend and enhance—i.e., *complement*—the capabilities of self-service front-end and data preparation tools. For example, the financial institution's data scientists and business analysts use a combination of data visualization and data preparation front-end tools to profile and prepare their data sets for analysis. A sizable proportion of the data they want to work with is already in the company's Teradata Data Warehouse. Previously, they had maintained a separate BI-analytic sandbox to support a self-service user experience for its data scientists and business analysts. Users would perform multiple extracts of data—in many cases, of the *same* data—from the Teradata Database to the BI-analytic sandbox. Data Lab simplified this process by eliminating time and resource-consuming data extracts. And because Data Lab is a collaborative environment, analysts can easily share their research with others—be they colleagues on the same team or co-workers in other business units. Thanks to Data Lab, the financial institution was able to compress the analytic lifecycle by as much as 66 percent, primarily because users are able to access, prepare, and process data much more quickly. Of equal importance is the fact that analysts can share and compare the results of their own analyses with colleagues and co-workers. To reiterate, this is one of the key strengths of the Data Lab model: data is processed *in situ*—i.e., in the context of the Teradata Data Warehouse itself. Data doesn't have to be extracted and moved to an external environment before it can be processed. All users of Data Lab are working and collaborating in the context of the Teradata system, against the same production data.

**GlaxoSmithKline** (GSK), a major global pharmaceutical company, tells of a similar experience . Before it switched over to Data Lab, GSK's analysts spent anywhere from 70 to 80 percent of their time preparing or "wrangling" data prior to the analysis. The shift to Data Lab permitted GSK's analysts to accelerate the rate at which data can be prepared and made available. In the first place, analysts can select, process, and access critical data in the Teradata data warehouse itself. It doesn't have to be moved to an external sandbox environment to undergo additional processing. In the second place, the MPP processing power of the Data Lab accelerates the rate at which data can be prepared, joined with production data, and analyzed. In sum, the shift to Data Lab permitted GSK's analysts to iterate much faster. It is now able to practice what Brad Donovan, a data analytics, informatics, and innovation leader with GSK, calls a "rapid prototyping mentality … allowed us to learn from an analytic exercise. If it wasn't working, we could scrap it and move on to the next [one]."

GSK's experience underscores the difference-making potential of MPP in-database processing: in one case, the time it took to process data decreased from 17 minutes (inclusive of data movement) in the external sandbox model to 14 seconds in the Data Lab environment. The faster analysts can process data, the faster they can iterate. Think of it as a "fail fast" process that enables analysts to eliminate what doesn't work and to quickly focus their efforts on what does. "A typical job that we might do would consist of data aggregating, model execution, model fitting, and quality control. We were running in hours, if not days, for some of this work, and by lifting and shifting a lot of this work into the Data Lab environment and leveraging the in-database capabilities of core analytics, we saw huge improvements," said Donovan. "We're going from 130 hours down to five hours in one example. We saw a huge uptick in productivity from the analyst community. [Data Lab] allowed them to move from a lot of the [unproductive] data wrangling exercises into more technical and strategic work."

Another Data Lab user, **Symantec® Corp.**, describes a similar experience . Symantec's users routinely pulled data from the Teradata Data Warehouse to populate dozens of rogue "data marts"—many of which consisted of desktop or laptop computers running Access databases or Microsoft SQL Server. Prior to Data Lab, Symantec's data warehouse team was powerless to police this practice: it couldn't offer data scientists and analysts a governable, manageable, scalable alternative that could support the agile methodology use case. Not only were these data marts ungoverned and unmanaged, but the process of populating them was slow and wasteful. (In a now-familiar trope, different users would often pull the same data for use in different data marts, resulting in multiple, redundant extracts.) Since implementing Data Lab, the data warehouse team has eliminated a sizable proportion of Symantec's rogue data marts. Data Lab has also helped accelerate the rate at which analysts and data scientists are able to develop and process their workloads. Finally, the switch to Data Lab has improved relations between the data warehouse team and the lines of business, says Robert Dissington, a senior IT architect with Symantec. "By giving the business ownership of that data lab, it reduces the time of delivery massively. You're no longer having to seek approval from IT, you're no longer having to get within the IT roadmap," he said. "You get a chance to rapidly prototype to test these hypotheses and see if it works."

> "The reduction of rogue data marts has been a very big goal of ours. We call it Shadow IT because we have no idea what's out there. The other thing… that's very nice is that we're leveraging server-level performance for this testing, [so] instead of having somebody's laptop running… a cut-down version of SQL Server, you're leveraging Teradata's MPP performance."
>
> – Robert Dissington, a senior IT architect with Symantec

Virtually all users of Data Lab interviewed relate some version of this story. They say that Data Lab gives their self-service users the ability to access, provision, and process data easily from the production data warehouse. They say Data Lab comprises a cost-effective and governable alternative to the traditional costly and siloed sandbox environment. They say Data Lab can eliminate the anarchy of the desktop or laptop sandbox. They say Data Lab isn't just a convenience, either: a majority of the organizations sampled said that their use of Teradata Data Lab permitted them to significantly compress their analytic development cycles.

Self-service users aren't the only beneficiaries of Teradata Data Lab. In most cases, data warehouse and BI programmers wind up using Data Lab in one way or another, too. Popular use cases include leveraging Data Lab as a prototyping environment for ETL and BI development.

For example, the online retailer uses Data Lab to accelerate its BI development cycle. Like most adopters, the retailer first deployed Data Lab as an agile analytic environment for its data scientists and analysts. But it also uses Data Lab as a prototyping environment. The idea is that once an analyst or data scientist identifies and refines an insight, she's able to prototype and test it—using production data—in Data Lab. The online retailer's data warehouse and BI teams use Data Lab in a similar way—i.e., as a test-dev prototyping environment– to accelerate conventional BI development. In its case, the online retailer's compressed its BI development cycle from almost two months—inclusive of post-development re-design and refactoring—down to three weeks or less.

The senior director of architecture with the advertising buying and selling arm of a prominent North American telecommunications carrier put it most succinctly. In this customer's case, too, Data Lab is used by both self-service analysts and data warehouse/BI programmers.

"Having Data Lab in the production environment enables you to experiment with whatever you're introducing, whether it's a new data source, or just another way to transform the data, how it's going to work or how it's going to perform," said the company's senior director of architecture.

## Capturing, Reusing, and Operationalizing Analytical Insights

Teradata Data Lab is able to support a viable and sustainable agile analytics user experience.

On the one hand, Data Lab provides an environment in which data scientists and business analysts can access , wrangle , collaborate, and, lastly, analyze data. This is Agility 101. However, as an analytic sandbox that *lives* in the Teradata system, Data Lab inherits the rich features and services of the Teradata ecosystem.

A viable and sustainable self-service practice must address the analytic lifecycle, from prototyping to production to ongoing maintenance to obsolescence. By contrast, the conventional analytic sandbox is a siloed environment. It provides few, if any, resources or mechanisms to reuse and disseminate information and insights. Ultimately, the capacity to productionize analytical insights is what makes an agile methodology viable and sustainable. It isn't enough to develop, test, and validate analytic insights; it is essential to *productionize* them, too. But prototypes don't always make it into production. Nor should they. For example, a niche or one-off use-case might make sense to maintain as a prototype. In the standalone sandbox model, this would be tantamount to the creation of still another rogue data mart. But the Data Lab model makes it easier to control and contain rogue data marts and spreadmarts. It standardizes and rationalizes the process of developing, testing, and validating prototypes. For these and other use cases, Teradata Data Lab provides governance and security features that can prevent disasters.

Switching to Data Lab permitted **GSK**, for example, to move away from a non-collaborative, siloed analytical development model. "We were a customer-facing analytics team that was embedded in SAS coding technologies, we were [focused on] creating insights and porting them out to Excel or PowerPoint. This was our core competency, but it didn't necessarily translate well to broader deployment in the organization," Donovan said. In the Data Lab environment, data scientists and analysts have the ability to test and train their models and algorithms against production data. This accelerates the development process and encourages a shift in thinking. Users begin to prioritize the development of production-ready analytics–in part, Donovan argues, because the Teradata environment simplifies the process of productionizing these insights.

The large North American financial institution uses Teradata Data Lab to power its self-service analytic practice. But it also uses Data Lab as a general-purpose prototyping environment in which to develop, test, and refine its BI and data warehouse projects. This has enabled the financial institution to accelerate the rate at which it develops and promotes projects into production. Like most large organizations, this financial institution has a disciplined software development lifecycle (SLDC). In most cases, a project must progress through each of the stages of this lifecycle (e.g., requirements analysis, development, testing, and quality assurance, among others) before it's promoted into production. Based on its experience with Data Lab, the financial institution implemented a policy that expedites Data Lab projects into production. The idea is simple: once a data scientist, analyst, or programming team tests and validates the value of a project in Data Lab, that project is rapidly productionized.

## Governance with an Eye towards Agility

In promoting a self-service experience, it's important to balance the rules of governance over and against the needs and priorities of users. In certain situations, for example, the rules and strictures of governance can and should be relaxed. Data quality levels are a great example. An organization (or the teams, groups, or business units that own a particular Data Lab) might set a baseline threshold for data quality—e.g., a data set must be no less than 66 percent accurate if it is to be used in analysis. There are bound to be cases in which a data scientist or a business analyst will want to run an analysis on data that is of much poorer quality, perhaps only 25 percent accurate. In such cases, a good argument can be made that even poor quality data is better than no data at all. Often, data scientists don't have any alternatives.

A viable and sustainable self-service experience must accommodate these and other scenarios. It must, therefore, have the capacity to be governed—or, depending on circumstances, *not* to be governed. Some governance, as with security or personally identifiable data, is mandatory at all times. But if the rules or mechanisms of governance are too strict, interest in sandboxes and discovery zones will atrophy as users feel stymied. They will look for alternatives. In the same way, if governance mechanisms are too relaxed, the lack of reasonable standards—e.g., the use of inconsistent (or improperly prepared) data sets or of incompatible/incomparable metrics— could have a no less pernicious effect. The result is that research, experimentation, and analytic development will stagnate, resulting in a critical diminishment in agility. When agility declines, the production of insights slows to a trickle. There are fewer insights to instantiate in the form of algorithms, rules, predictable data uses, and new business opportunities.

Interviews with customers demonstrate that the *capacity* to govern distinguishes the Data Lab model from that of the traditional sandbox environment, or, for that matter, the recent crop of do-it-yourself front-end tools. Unlike an analytic sandbox or self-service analytic tool, Teradata's approach permits the user or group that owns a Data Lab sandbox to enforce rules. Owners can impose access control restrictions, restrict access to potentially sensitive information, or implement mechanisms that can be triggered to mask potentially sensitive information. IT can provision a single large Data Lab environment for all internal users or, more commonly, a different Data Lab sandbox for each user, team, group, or unit. IT can also delegate the internal management and administration of each sandbox to the user, team, group, or unit that controls it. In this way, *users themselves* have the freedom to create, manage, and enforce the rules that govern the storage, use, archiving, and deletion of data inside of their own Data Lab environments.

Discipline and self-enforcement is required of user departments that utilize Data Lab. While Data Lab is safer and more governed than data-marts-under-my-desk, users also have the ability to misbehave or misuse the data. Distributed governance means distributed responsibility. Several users of Data Lab have actually created programs to educate users as to the importance of using data accurately, ethically, and responsibly.

> "What we found to be very successful was the creation of what we call the Data Lab Cookbook.... We worked with Teradata Professional Services to create a very GlaxoSmithKline-specific document that really encouraged us to collaborate, [to understand the importance of] security, governance, data management and tools. It simplifies the process of getting Data Lab up and running."
>
> – Brad Donovan, GlaxoSmithKline

The North American financial institution likewise took pains to educate its users about the importance of governance. In its case—i.e., as a financial institution–it has to be extremely careful about the balance it strikes between empowering Data Lab users and enforcing security and governance. Probably the biggest challenge in using Data Labs is balancing the needs of good governance over and against those of end-user agility. In the case of the financial institution, it needed to enforce policies in the Data Lab environment that align with its own governance policies. Depending on how much latitude IT gives to the lines of business, it's possible to override compliance settings in the Data Lab environment. For this reason, the financial institution implemented a training and education program for Data Lab, the emphasis of which is on making responsible and ethical use of data. To drive these points home, it warns Data Lab users that the onus for using data responsibly is on them.

## Conclusion

The Data Lab environment is not a panacea. On its own, it won't make or break the success of an agile environment. Now as always, the contributions of human actors—acting singly and/or in combination with one another—are absolutely critical, as are the processes they create, adhere to, or use to serve their own ends. Above all, an organization's culture—the gestalt of its people and processes—is the critical determinant. In a tighty controlled, tightly governed organization, agility simply cannot and will not thrive, with or without an agility-friendly environment like Data Lab; with or without self-service front-end tools; with or without the brightest and most imaginative data scientists, predictive modelers, business analysts, and data architects. This was the case with the advertising buying and selling arm of a prominent North American telecommunications carrier. This company's culture is by no means repressive, its analysts and its BI/data warehousing team are constrained by its overly restrictive SLDC governance. "We're being hamstrung now because it's so hard to put stuff into production. We're trying to play the game, but it's been hard to do this [i.e., to expedite analytical insights into production]," said the company's senior director of architecture. In contrast, the North American financial institution adapted its SLDC process to realize the benefits of rapid prototyping, of analytic insights in Data Lab.

For an organization that is committed to developing a viable and sustainable agile analytic practice, however, an analytic environment such as Data Lab is a must-have. This isn't in any sense to say that Data Lab is *the only* solution; rather, it's recognition that something *like* Data Lab is a necessary component of a viable and sustainable agile analytic practice. It is necessary to complement self-service front-end tools with an analytic environment that permits users to:

- Quickly access production data;
- Process data in parallel to maximize performance and accelerate iterative development;
- Provision their own data sets, feeds, and sources;
- Configure their own rules to govern and manage the environment;
- Reliably productionize analytic insights.

For existing Teradata customers, Data Lab is a no-brainer. It will eliminate costly, wasteful, and redundant external analytic sandboxes, and is also of value to the BI and data warehousing teams.

## About the Author

Stephen Swoyer is a technology writer with 20 years of experience. His writing has focused on business intelligence, data warehousing, and analytics for almost 15 years. Swoyer has an abiding interest in tech, but he's particularly intrigued by the thorny people and process problems technology vendors never, ever want to talk about.

## About Our Sponsor

Teradata helps companies get more value from data than any other company. Our big data analytic solutions and team of experts can help your company gain a sustainable competitive advantage with data.

## Endnotes

1. Webinar, Deploying Agile Analytics for Business Innovation, September 17, 2015, http://event.on24.com/wcc/r/1032794/DD207F01BE8E4583FE417A25313643D5
2. Webinar, Deploying Agile Analytics for Business Innovation, September 17, 2015, http://event.on24.com/wcc/r/1032794/DD207F01BE8E4583FE417A25313643D5

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