Paper 201

Data Labs with SAS® and Teradata: Value, Purpose and Best Practices

Bob Matsey Teradata Corporation, Charlotte, North Carolina

Tho Nguyen, Teradata Corporation, Raleigh, North Carolina

Paul Segal, Teradata Corporation, San Diego, California

ABSTRACT

Success in today's global marketplace requires the agility to quickly test hypotheses and experiment with new concepts that drive innovation, uncover market opportunities and proactively react to competitive pressures. Companies need the ability to extend their analytic practice with the proper infrastructures, tools and technology to enhance the agility of analyzing data. That means adding critical flexibility for exploring new, unrefined data, pulling in external data sources and experimenting on emerging theories with 'production' Integrated Data Warehouse (IDW) data without long planning periods or always having to get IT to load their data. By considering a data lab, its value and purpose is to explore and examine new ideas, experiment, prototype new possibilities by combining new data with existing IDW data to create experimental designs and ad-hoc queries without interrupting the production environment.

INTRODUCTION

Companies are collecting more data than ever before, and data marts, SAS datasets, Access databases, Excel spreadsheets are all proliferating throughout the enterprise, producing inconsistent analytic results and driving up IT cost with inefficient processes and redundant systems. To avoid these pitfalls, setting up a data lab has become a very popular and strategic option to enable a self-service, collaborative environment for analytic development - to test and experiment with the data without impacting the production data warehouse platform. Teradata has introduced the Data Lab concept to do just that for many years and it is also known as a 'play-pen' or 'sandbox' to work with untested data or to submit queries for experimental designs. The data lab environment enables business analysts & scientists to access data that previously wasn't easily available to join with production data and enables businesses to explore and experiment with new ideas and data. Data Labs makes it easy to identify new trends, develop new insight in your data or react to immediate business issues/opportunities.

This paper will cover the following topics:

- Purpose of the Teradata® Data Lab with SAS
- Value of the using SAS and Teradata
- Use cases and best practices

PURPOSE OF THE TERADATA ® DATA LAB WITH SAS

A Teradata Data Lab enables self-service 'Agile' analytics and BI by simplifying the provisioning and management of analytic workspace within the production data warehouse. By allocating that space in the Data Lab environment, it provides your lab users with easy access to critical production information without moving or replicating that data. And it provides your entire business with flexibility to self-provision space, load other types of data into the data lab that they can experiment with this new data and test theories and business ideas without compromising data quality or performance of the production data warehouse. The result is an environment that allows your business to maximize hardware, software and human resources; streamline your processes; and deliver fast and accurate answers while controlling the costs and other issues that stem from data mart proliferation and redundant data.

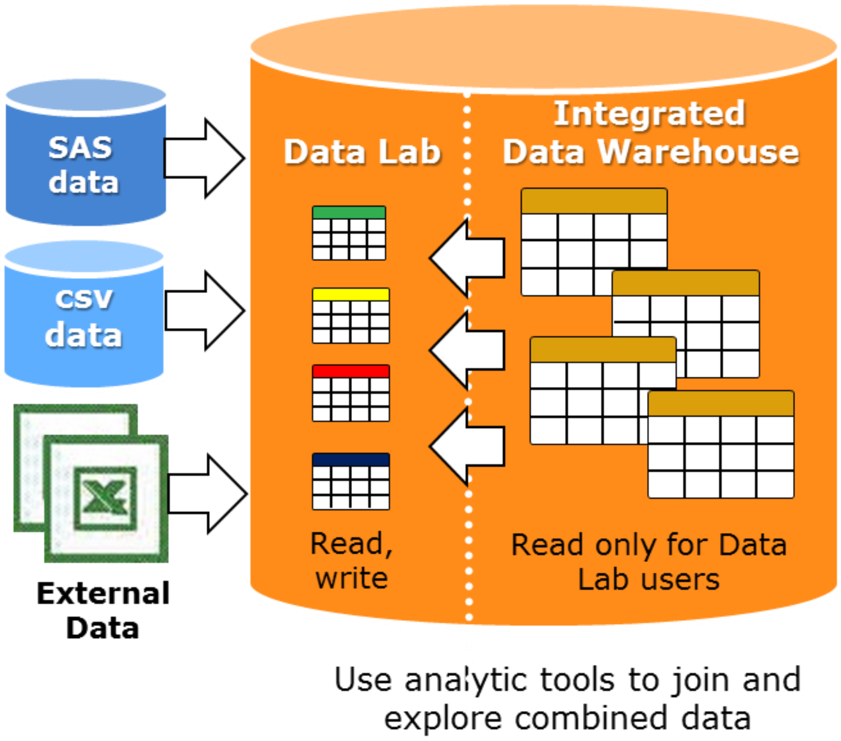

SAS data
csv data
External Data
Data Lab
Integrated Data Warehouse
Read, write
Read only for Data Lab users
Use analytic tools to join and explore combined data

Figure 1: Teradata® Data Lab with SAS®: Promote Experimentation and Exploration of Data

The Data Lab can be set up within your Teradata production warehouse. The data can be joined through views with the Integrated Data Warehouse with no data exports. One of the key concepts of the data lab is to 'self-load' new or experimental data quickly which can be used for rapid prototyping, experimentation and exploration analysis – ' Agile Analytics'. Key benefits to the business include: easy to use for self-provisioning and management, extending analytics to many more users and minimize IT support after initial setup.

VALUE OF THE DATA LAB WITH SAS AND TERADATA

As referenced in Figure 1, the data lab can be set up within the data warehouse environment and offers many advantages to both IT and business users.

For IT Users

Teradata Data Lab safeguards and controls production data to ensure predictable workload performance and segregation of the jobs and maintains data quality within the data warehouse by reducing data redundancy. It also provides an GUI interface to easily establish guidelines, timeframes and rules for the work area by controlling size, expiration dates and user access. Data Lab implements these guidelines to help business analysts manage individual work areas and also provides these advantages to your IT department:

- Control the environment with data lab hierarchy and workload management to ensure production users are not impacted by the exploratory nature of analytic development.
- Governance is provided by enforcing expiration dates, exception processes, with a Data Lab Governance board reviewing data lab activity, growth, additional needs while preventing unintentional use or growth of data labs.
- Eliminate shadow IT and personal external data marts/Access DB/Excel spreadsheet growth to help control costs.

- Optimize data integration by identifying and integrating only the data that provide high proven business value.
- Free's DBA's & IT from having to load experimental data and provision space on an ongoing basis

For Business Users

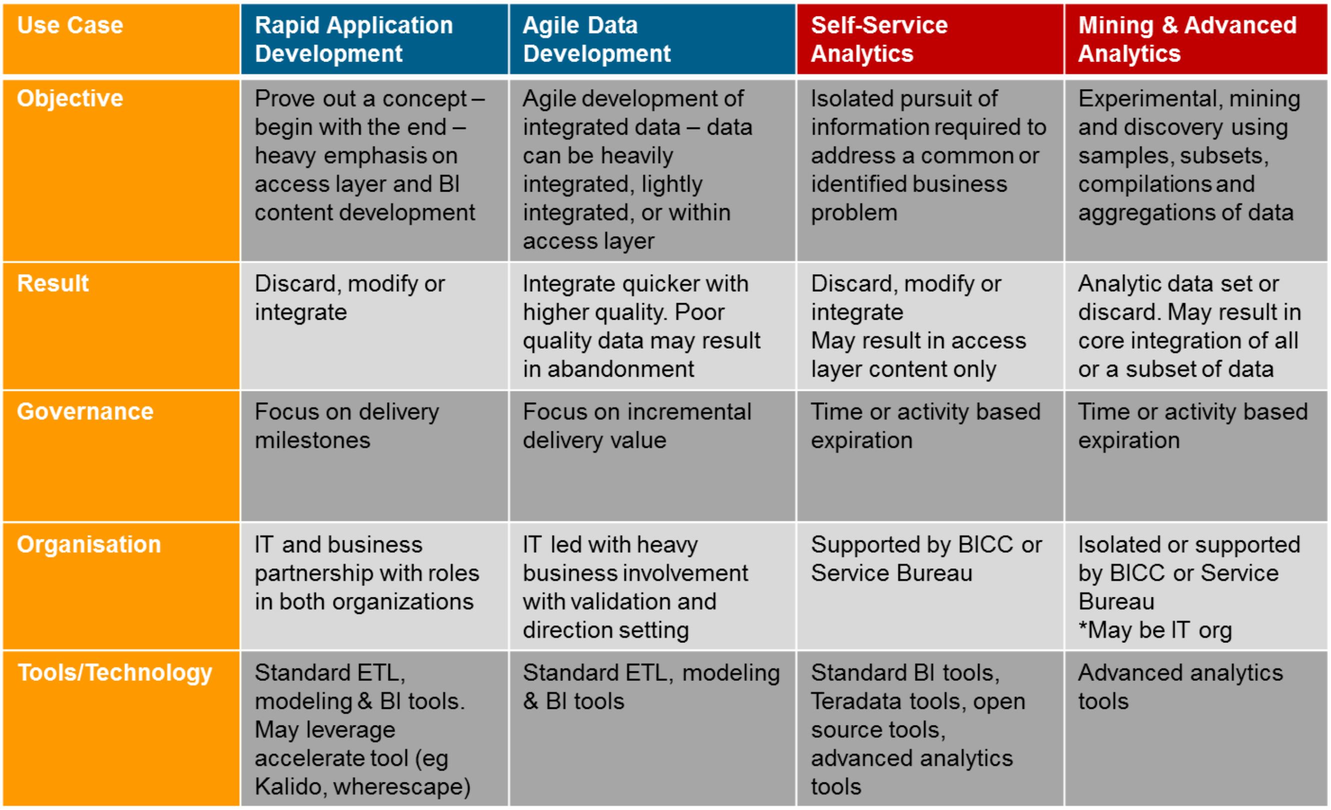
Teradata Data Lab provides an interface that allows your business users to easily provision space for exploration and experimentation of external data within their production data warehouse. This GUI interface allows business users to load and analyze data using their favorite tools or share their analysis and data with others authorized users. Teradata Data Lab simplifies collaboration, self-service and data access by eliminating unnecessary data movement and replication. Here are some of the advantages that Teradata Data Labs offers for your business users:

- A rapid exploration, experimentation and development environment with self-service analytics, provisioning and management capabilities.
- Quick time to value by accelerating and simplifying the process of loading and management for business users.
- Eliminate data movement & data replication by allowing users to load and join new data with production data.
- Access to new data sources through self-loading tools without waiting for a full data warehouse integration project.
- Simplified collaboration: Data lab users can easily invite others to share their analytic data and results based on proper security authorization.

USE CASES AND BEST PRACTICES

There are multiple factors in play that influences the ultimate design and governance of the *Data Lab*. The first factor is where the customer is in regards to an EDW maturity lifecycle. In the scenario where the Teradata Customer is new to Teradata, an approach that applies soft or no governance initially is acceptable, as it is more important to get the *Data Lab* space in place before the business users begin creating it on their own through various means, such as spreadsheets, data marts or SAS datasets. The new Teradata Customer is typically driving hard towards realizing business value immediately, rather than awaiting full data warehouse population. For more mature Teradata Customers, before the transition from legacy sandboxing into *Data Labs* it is imperative that a governance model be defined and set up to ensure proper process, procedures and controls for the environment. This model should provide the structure for utilization/request process of the *Data Lab* and exception and exit processes, including any industrialisation that must occur at the end of their duration. In the *Data Lab* context, we will refer to the development of a standardized, sustainable method of integration into the Teradata Customer's enterprise standards for migration to production. As part of a transition from sandboxing into *Data Lab,* the Teradata Customer has the opportunity to restart the program, under a new name with new structure to optimize the activity in the environment. Foregoing this governance design as a new user early may result in difficulty in establishing such a model at a later date due to it's very difficult to take something away from some who's had no timeframes or durations set initially in the data lab environment set up.

The *Data Lab* architecture is leveraged through multiple use cases, designed to support specific business functions. At the highest level, there are two main use case categories that are aligned specifically with the business users control and function for each. There are fundamental similarities across all use cases, such as simple provisioning, flexible data sourcing, including table creation, and pre-defined governance of its utilization and duration period. The variations include who performs the hands-on roles of the data sourcing and table design between business and IT. Regardless of the use case, the overall objective remains the same: faster data access for analysis, validation and reporting needs.



Use Case	Rapid Application Development	Agile Data Development	Self-Service Analytics	Mining & Advanced Analytics
Objective	Prove out a concept – begin with the end – heavy emphasis on access layer and BI content development	Agile development of integrated data – data can be heavily integrated, lightly integrated, or within access layer	Isolated pursuit of information required to address a common or identified business problem	Experimental, mining and discovery using samples, subsets, compilations and aggregations of data
Result	Discard, modify or integrate	Integrate quicker with higher quality. Poor quality data may result in abandonment	Discard, modify or integrate May result in access layer content only	Analytic data set or discard. May result in core integration of all or a subset of data
Governance	Focus on delivery milestones	Focus on incremental delivery value	Time or activity based expiration	Time or activity based expiration
Organisation	IT and business partnership with roles in both organizations	IT led with heavy business involvement with validation and direction setting	Supported by BICC or Service Bureau	Isolated or supported by BICC or Service Bureau *May be IT org
Tools/Technology	Standard ETL, modeling & BI tools. May leverage accelerate tool (eg Kalido, wherescape)	Standard ETL, modeling & BI tools	Standard BI tools, Teradata tools, open source tools, advanced analytics tools	Advanced analytics tools

Figure 4 – Data Lab use cases

Agile Development

An agile architecture, utilizing a *Data Lab*, provides the foundation for leveraging an agile, rapid application development methodology. This approach can be utilized with a formal, Agile Methodology, such as Scrum or XP, Teradata's PS offering for Data Lab establishment, or merely through a process or practice change within the Teradata Customer's organization. In this category, there are the following use cases:

- Rapid Application Development/Prototyping

In this use case, which can include variations of load-and-go analytics along with Proof of Concept (POC) applications, the business user's role is that of requirements and analysis provider, data consumer and data validation. The IT/Data Warehouse team typically is the builder of the data, in multiple iterations with shorter delivery cycles, aimed at providing the business a quicker time to value/time to failure process. These IT/Data Warehouse team activities may be performed by the data sourcing/integration team, or by a **BI Competency Centre (BICC)** or **Service Bureau** team**,** (referred to as *The Centre* going forward).

With a prototype, the business user can realize the value of data or an application and can be able to see what they will get after industrialisation. They are able to define their requirements at a very high level of detail without any misunderstandings. In case of urgency, the prototype can deliver information in the interim, while the industrialization process is being developed. After the industrialization is finished, the users only have to "switch" to the production area. Also the prototype can be able to save a significant amount of money if the business user sees that the value of the data, reports or analytics of the prototype is too low and decide that it should not be implemented. Instead, they are now able to turn their focus to another project that does demonstrate the appropriate value.

- Agile Data Development

In the Data Development use case, the goal starts with and ends with a new source of data. The activities that occur in the *Data Lab* include, but not limited to, evaluating the usefulness of the data, data quality profiling and the definition of the integration points of the data for use with other data across the data warehouse.

Business User Self-Service Analytics

Business User Self-Service Analytics are characterized by a hands-on control of sourcing and consuming the data by the business user, broken into multiple use cases, including the following:

- Self-Service Business Usage

This use case is typically subdivided by business unit, although it can be designed and governed in smaller increments, as little as a single user. The usage scenarios include business unit specific aggregates or views, hierarchy tables to facilitate reporting, external 3rd party data acquisition and evaluation, and sourcing of internal data currently not contained within the data warehouse. The *Data Lab* provides the business users an area for analysing and exploring new and already available data in the data warehouse. The business users do have full freedom to do data research in order to determine/define the value of different kinds of data transformation, analysis, reports or the modification of transformations that may be required. It is expected that in many of the activities the IT organisation team will not be involved. Activities of the users within their own *Data Lab* can be, however, supported by *The Centre*. These activities may include the loading of data or building of complex SQL.

- Advanced Analytics

This use case utilizes data mining, predictive analytics, applied analytics, statistics and other approaches driven from techniques that help develop models, new views of data and simulations to create scenarios, understand realities and predict future states. With the supporting *Data Lab* architecture in place, the traditional barriers to action are removed or minimized. No longer does the business user have to wait for space to be allocated, data to be sourced and access to be granted to begin the act of developing such models to make immediate business impact.

In some cases, an application built in this context may be lasting (A Score Factory, for example). Particular care must be taken over whether this lasting mode is suitable for use in the *Data Lab*, as it is not intended to become a substitute production environment. The appropriate approach should be defined as part of the *Data Lab* governance program.

Identifying the use case that applies

When designing the *Data Lab* architecture, the foundation requirements will be similar across the various use cases, enabling re-use of the architecture for multiple scenarios. The governance around the space, including the owner of creation, sourcing and process development is a key variable. As discussed above in the high level use cases, the business user’s role is a key element into how the environment is managed, from self-provisioning to data access to data retention and expiration.

Typically most Teradata Customers will identify multiple use cases, or begin with a pilot use case, and evolve the *Data Lab* environment to support other initiatives while tweaking the Governance, process and procedures to their need. The use cases above are just examples, and few customer use cases fit into just one of these with most having parts of 1 or many of them. As an example, it may be best to enable self-service analytics for the business user, and, after they become comfortable with environment and tools, expand the usage to one of the data development use cases, to involve the business user earlier in the development cycles via the *Data Lab*.

CONCLUSION

Teradata Data Labs provides the foundation for self-service Agile analytics and business intelligence by providing an interface to automate much of the provisioning and management of an analytic workspace. And it offers the on-demand capacity you need for short-term projects such as ad-hoc analysis, data mining, proof of concept testing and quality assurance testing.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Bob Matsey
Enterprise: Teradata
E-mail: bob.matsey@teradata.com

Name: Tho Nguyen
Enterprise: Teradata
E-mail: tho.nguyen@teradata.com
Web: www.teradata.com/sas

Name: Paul Segal
Enterprise: Teradata
E-mail: paul.segal@teradata.com