# Conquering Big Data Analytics with SAS, Teradata and Hadoop

John Cunningham, Teradata Corporation, Danville, California

Tho Nguyen, Teradata Corporation, Raleigh, North Carolina

Paul Segal, Teradata Corporation, San Diego, California

## ABSTRACT

Organizations are faced with the unique big data challenges collecting more data than ever before, both structured and unstructured data. There has never been a greater need for proactive and agile strategies to overcome these struggles in a volatile and competitive economy. Together, SAS and Teradata have joined forces with Hadoop to revolutionize your business by providing enterprise analytics in a harmonious data management platform to deliver strategic insights. This paper discusses how SAS, Teradata with Hadoop are delivering innovation to break through your big data analytics challenges, by exploring the appropriate platforms for the various types of analysis to quickly uncover hidden opportunities.

## INTRODUCTION

Big data is often defined by the 3 Vs: variety, velocity and volume. There is a fourth dimension and it is value. Variety implies to the different types of data, formats and patterns which can be structured and unstructured (or semi-structured). Velocity is the speed of data collection, consumption and analysis of data in a timely manner. Volume is associated with size of the data and files and has the most impact. Many organizations treat data as a strategic asset and organizations are collecting more data than ever before for historical analysis. Big data presents great opportunities but also challenges to analyze ALL of that complex data and ultimately derive value. Value can be achieved with the appropriate architecture and technology.

To overcome these challenges and obtain value, we introduce some innovative approaches to managing and analyzing big data along with an architecture that is unified for analytics and data management. The architecture showcases the integrated technologies from SAS, Teradata and Hadoop, allowing organizations to effectively manage big data and apply the analytics directly to the data to quickly derive value for effective decision making and competitive advantage.

This paper will cover the following topics:

- SAS® Analytics for Teradata
  - In-database analytics
  - In-memory analytics
- Teradata Appliance for SAS® High-Performance Analytics, Model 720
- Hadoop in the data architecture
- Teradata® Unified Data Architecture$^{TM}$
- Bringing it all together

## SAS® ANALYTICS FOR TERADATA

For the past seven years, SAS and Teradata have delivered a number of programs and offers integrating SAS analytics inside the Teradata family platform. The intent of these programs and joint offers is to provide customers solutions that reduce the complexity managing big data analytics and cost for effective decision making. Together, SAS and Teradata have joined forces to deliver innovations by integrating the best of breeds and combining analytics and data management in a unified solution. Our solutions offer end-to-end capabilities ranging from data exploration, data preparation, model development and model deployment. We have developed horizontal and vertical offers to meet customers' needs specifically for big data analytics.

There are two key technologies that dramatically improve and increase performance when analyzing big data: "in"-database and "in"-memory analytics.

## IN-DATABASE ANALYTICS

In-database analytics refer to the integration of advanced analytics into the data warehousing. With this capability, analytic processing is optimized, to run where the data reside, in parallel, without having to copy or move the data for analysis. Many analytical computing solutions and large databases use this technology because it provides significant performance improvements over the traditional methods. Thus, in-database analytics have been adopted by many SAS business analysts who have been able to realize the benefits of streamlined processing and increased performance. With SAS® in-database analytics for Teradata, SAS users have the ability to develop complex data models and score the model in the data warehouse. By doing so, it removes the need to either move or extract the data to a SAS environment or convert the analytical code to something that could be executed on the data platform.

By applying the analytics to where the data reside, it significantly streamlines the process by eliminating data movement and redundancy. In addition, it greatly improves data integrity by not having to copy and move the data to a silo data server. The improved performance comes from leveraging the power of the Teradata data warehouse with its massively parallelize processing (MPP) architecture. The MPP architecture is a "shared nothing" environment and can take disseminate large queries across nodes for simultaneous processing. It is capable of high data consumption rates through parallelized data movement which means completing any task at a fraction of the time. The diagram below illustrates the in-database processing.

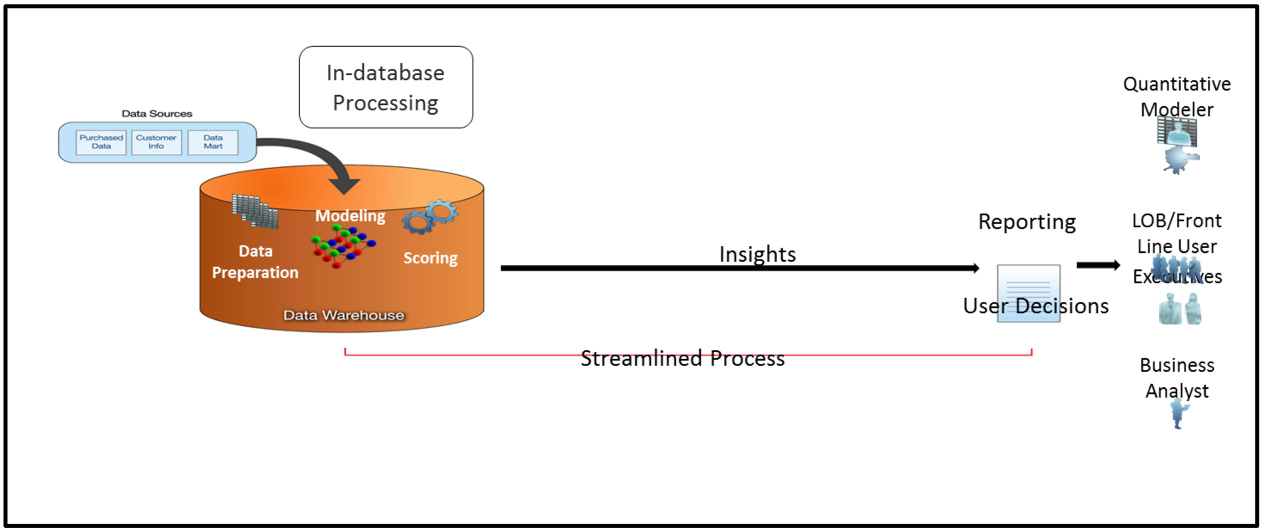

**Figure 1**: In-database processing: Minimize data movement and redundancy

In-database processing includes data preparation, data modeling and model scoring – all of which can be executed inside the Teradata data warehouse. The in-database approach dramatically streamlines the process compared to the traditional method and insights can be delivered to business and IT faster for informed business decisions.

As referenced in Figure 1, data preparation can be executed inside the data warehouse. For data preparation, the following products are integrated with Teradata

- SAS/ACCESS® Interface to Teradata - a data adapter that can interface directly with Teradata
- BASE SAS – a selected set of PROCS -
    - PROC SUMMARY
    - PROC MEANS
    - PROC FREQ
    - PROC RANK
    - PROC TABULATE
    - PROC REPORT
    - PROC SORT
- SAS Data Quality Accelerator for Teradata – data quality functions to cleanse and integrate the data
    - Matching
    - Parsing
    - Extraction
    - Standardization
    - Casing
    - Pattern analysis
    - Identification analysis
    - Gender analysis
- SAS Code Accelerator for Teradata - simplifies and speeds data preparation with user-defined methods utilizing DS2 programming language

For data modeling, the following products are integrated with Teradata

- SAS Analytics Accelerator for Teradata – a set of PROCs to develop and deploy models
    - SAS/STAT
        - PROC REG
        - PROC PRINCOMP
        - PROC VARCLUS
        - PROC SCORE
        - PROC CORR
        - PROC FACTOR
        - PROC CANCORR
    - SAS Enterprise Miner
        - PROC DMDB
        - PROC DMINE
        - PROC DMREG (Logistic Regression)
    - SAS ETS
        - PROC TIMESERIES

For model scoring, there following products are integrated with Teradata.

- SAS Scoring Accelerator for Teradata – scoring of models from SAS Enterprise Miner and SAS STAT

In addition to the above products and capabilities, we have additional in-database offers and solutions.

- **Business Insight Advantage Program** - A complete certified solution for Data Management & Quality, Business Intelligence and Analytics that includes Teradata Database & hardware, SAS software and joint services

- **Anti-Money Laundering (AML) Advantage Program** – A complete Anti-Money Laundering solution built around SAS AML with Teradata for running scenarios and risk factors in-database.

- **Credit Risk Advantage Program** - A solution integrating SAS Credit Risk with Teradata and the Financial Services Logical Data Model FS-LDM.

- **Credit Scoring Advantage Program** - Execute SAS Credit Scoring functions inside the Teradata database at extraordinary speed to manage credit application adjudication and portfolio management.

- **Warranty Analysis Advantage Program** – A combined solution of SAS Warranty Analysis and Teradata Early Warning Analytics with associated hardware and software, services.

As the partnership matures, we have evolved from in-database to in-memory analytics.

## IN-MEMORY ANALYTICS

The SAS in-memory environment leverages Teradata's MPP (Massively Parallel Processing) architecture which is ideal for retaining, preparing and partitioning large data sets for big data analytics. It is capable of high data consumption rates through parallelized data movement which means completing any task at a fraction of the time. This latest innovation provides an entirely new approach to tackle big data by using an in-memory analytics engine to deliver super-fast responses to complex analytical problems. It is a set of products beyond SAS Foundation technologies to explore and develop data models using all of your data.

The SAS Foundation software is located on a user's workstation or on a SAS server. When it runs a SAS program containing High-Performance procedures or analytics, it initially connects to the Teradata database containing the source data, and then it instigates a parallel computing job on the SAS processing nodes. One of the SAS nodes is designated to be the controlling root node and the other nodes are worker nodes.

The SAS client coordinates with the root node, and the root node in turn directs with the corresponding processes on the worker nodes. The worker processes are multi-threaded to take advantage of the large number of CPUs. Therefore, once an in-memory analytics process runs on the appliance, all of the nodes are dedicated to that specific task. Analysis can be executed in minutes or seconds using this approach.

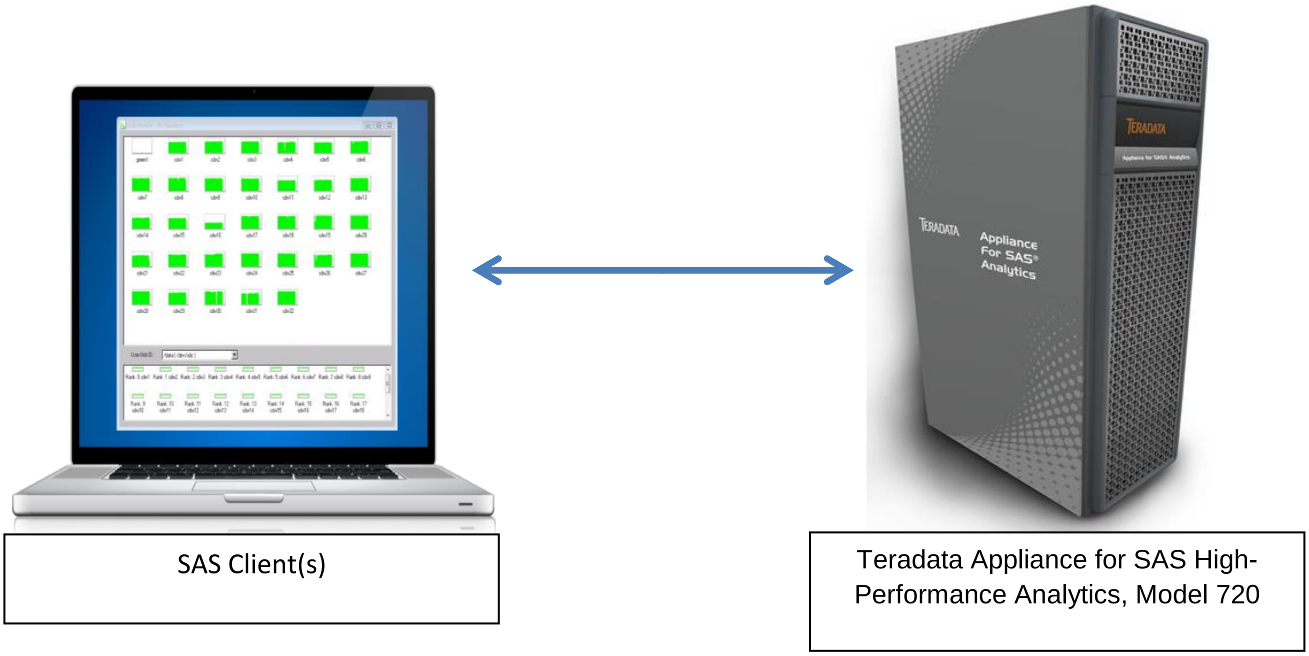
TERADATA
Appliance For SAS® Analytics
SAS Client(s)
Teradata Appliance for SAS High-Performance Analytics, Model 720

**Figure 2**: In-memory processing

When all of the processes are running for an in-memory task, the root node submits a SQL query to Teradata that causes the SAS Embedded Process (EP) table function to read data from the database and send it to a SAS in-memory worker. Teradata was designed to multi-threaded. For a specific SQL request, each thread is called an AMP worker thread. Since the SAS EP is also multi-threaded, it makes a connection from every Teradata AMP to a SAS worker.

After the data is transferred to memory and while the SAS in-memory job is active, there is no activity in the Teradata database. Thus, there is no performance impact to the Teradata database as data is only lifted into memory when requested.  SAS software coordinates the analytical processing between the SAS client that is running the procedure, the SAS HPA root node, and the SAS worker nodes. All of the nodes in the Teradata Appliance for SAS are designated to compute the analytical tasks.

When the SAS HPA in-memory processing is complete, results can be written back to Teradata into a permanent client for additional analysis, depending on the type of procedure and the procedure options that are selected.

## TERADATA APPLIANCE FOR SAS HIGH-PERFORMANCE ANALYTICS, MODEL 720

The Teradata® Appliance for SAS High-Performance Analytics, Model 720 is specifically for SAS High-Performance Analytics Products and SAS® Visual Analytics, integrating SAS in-memory capabilities with the industry leading data warehouse platform, for data model development and data visualization. Jointly developed with SAS, the Teradata Appliance for SAS High-Performance Analytics, Model 720 eliminates the need to copy data to a separate appliance with dedicated SAS nodes for in-memory processing.

There are a number of SAS products that seamlessly integrate with the Model 720.

- **SAS Visual Analytics** - Explore massive volumes of data to quickly to visualize and uncover patterns and trends for further analysis
- **SAS High-Performance Analytics Products**
    - **SAS High-Performance Statistics**: Enables use of predictive models for faster and more effective decision-making.
    - **SAS High-Performance Data Mining**: Develops predictive models using thousands of variables to produce more accurate and timely insights.
    - **SAS High-Performance Text Mining**: Explores all your data, including textual information, to gain rich new knowledge from previously unknown themes and connections.
    - **SAS High-Performance Forecasting**: Generates models for faster high-value and time-sensitive decision making, using thousands or even millions of granular-level forecasts.
    - **SAS High-Performance Econometrics**: Provides econometric modeling facility, such as the number and severity of events, using big data.
    - **SAS High-Performance Optimization**: Performs more frequent modeling iterations and uses sophisticated analytics to get answers to questions you never thought of or had time to ask.

By leveraging analytical features, including statistics, data mining, text mining, forecasting econometrics and optimization, organizations can quickly identify and add important variables. More data model iterations can be performed to gain understanding and make decisions with confidence.

The Teradata Appliance for SAS High-Performance Analytics readily extends the entire Teradata Platform Family as shown in Figure 3, providing ultra-high speed SAS® In-Memory Analytics against Teradata Data Warehouses and Appliances. The appliance features clustered servers, each with dual Intel® eight core Sandy Bridge processors, SUSE® Linux operating system, 128-256GB of RAM, and enterprise class Infiniband networking infrastructure—into a power-efficient system. The appliance connects directly to Teradata BYNET, ensuring unsurpassed data access speeds, 50-250x faster than traditional ODBC, and superior analytic processing. Best of all, the solution is supported by the most trusted name in data warehousing—Teradata.

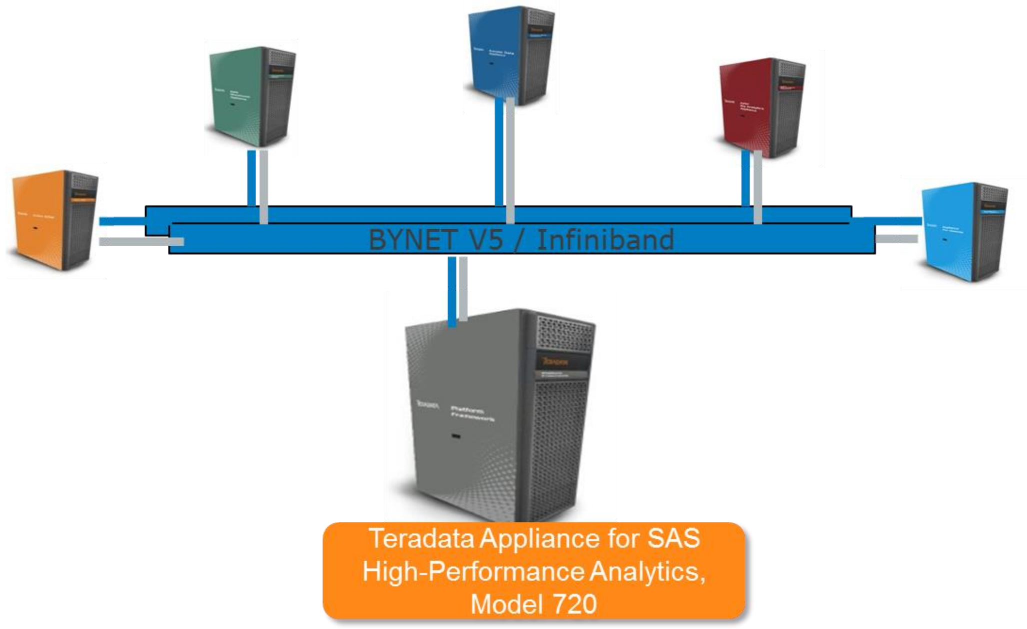
BYNET V5 / Infiniband
Teradata Appliance for SAS High-Performance Analytics, Model 720

**Figure 3:** Teradata Platform Family Connects with Model 720

The Teradata Appliance for SAS High-Performance Analytics, Model 720 enables advanced analytics with incredibly fast parallel processing, scalability to process massive volumes of data, and rich in-memory analytics capabilities. This environment provides a set of in-memory analytics algorithms that leverages the database's speed, while eliminating time-consuming and costly data analysis. This Teradata appliance includes analytical capabilities spanning data visualization and data model development executed in a highly scalable, in-memory processing architecture. It will let customers explore massive volumes of data with SAS Visual Analytics and develop analytical models using complete data—not just a subset—with SAS High-Performance Analytics products to get accurate and timely insights and make well-informed decisions. Often faced with hundreds of candidate variables, this offering helps to determine unimportant variables, describe important relationships, and identify the important factors for subsequent models and data exploration.

The Teradata Appliance for SAS High-Performance Analytics is easy to manage, you can free up your DBA resources to do other valuable tasks. With virtualized CPU, memory, and storage all designed to work together as a unit, you get automated management of physical disk space so your DBAs never have to worry about data placement or data reorganization.

With this Teradata appliance, companies can start with a small configuration, and then expand as needed driven by the ongoing analytic needs of the business.

## HADOOP IN THE DATA ARCHITECTURE

Every business has it uniqueness – with different needs, requirements and architecture. Hadoop has emerged in some businesses that have taken advantage of the inexpensive storage on commodity servers, complementing the existing data warehouse, discovering platform, business intelligence and data management systems. With many big data analytics projects, organizations are exploring and adopting Hadoop in their data architecture to support the new platform known as a "Data Lake." This flood of data, ranging from diverse formats to new data sources that were not formerly collected, is now driving the need for a modern platform called a "Data Lake," where data can economically and efficiently be captured, stored and refined to support analytics.  This phenomenon introduces enormous challenges managing, analyzing and exploring the petabytes of structured and semi-structured data without making redundant data and analytics scattered around the company.

The interest in Hadoop is frequently associated with highly specialized business problems. Although it may be the compelling to make a redundant copy of the data, perform value added analytic services, and potentially provide a high value analytic answer, it may not be the best and most effective approach in the long run. The challenges include:

- Copying data onto a disconnected Hadoop cluster can be a slow, tedious process. Worse yet, copy high value answers back to some other operational system for tightly integrated services can be equally as painful. If you're a highly valued data scientist, you may spent a large amount of your time waiting for large datasets to be copied – just to get started on the analytic work that matters.
- If your experiment is successful, once you spin one of these dependent Hadoop clusters up, the cluster needs to be managed. That likely requires ongoing Hadoop administration, managed ETL from disparate data sources, accessibility control, etc.
- Also, by definition, once you copy data to a redundant data mart, it's redundant! This means that the same data is in two different places, and is being transformed differently in each – which ultimately will lead to varying results and lack of data governance.

For the challenges above, we highly recommend an integrated and unified architecture.

## TERADATA UNIFIED DATA ARCHITECTURE – PURPOSE BUILT DATA MANAGEMENT

Teradata has long been the industry leader in large scale Enterprise Data Warehouse (EDW) platform, and most of the biggest companies, with the biggest data warehouses, are running on Teradata. The Teradata Unified Data Architecture is an evolution of the model, and recognizes that an EDW isn't always the right place for certain classes of data. For example, while Teradata Data Warehouse scales well to support complex data models, access from hundreds of users, with varying workloads, in a highly managed, highly governed, rigorously controlled environment, there are classes of data that are likely not going to benefit from Teradata services.

In addition, while Teradata may support certain unstructured data, log files, or streaming sensor outputs to be written into the Teradata database in the form of Binary Large Objects, putting them there may not provide any analytic lift to the data scientist. Data types like BLOBS are very capable of store virtually any random data content, but if the end goal is to support parsing, text analytics, keyword lookups, then other stored methods may be preferable.

For that reason, we introduce the Teradata Unified Data Architecture as shown in Figure 4.

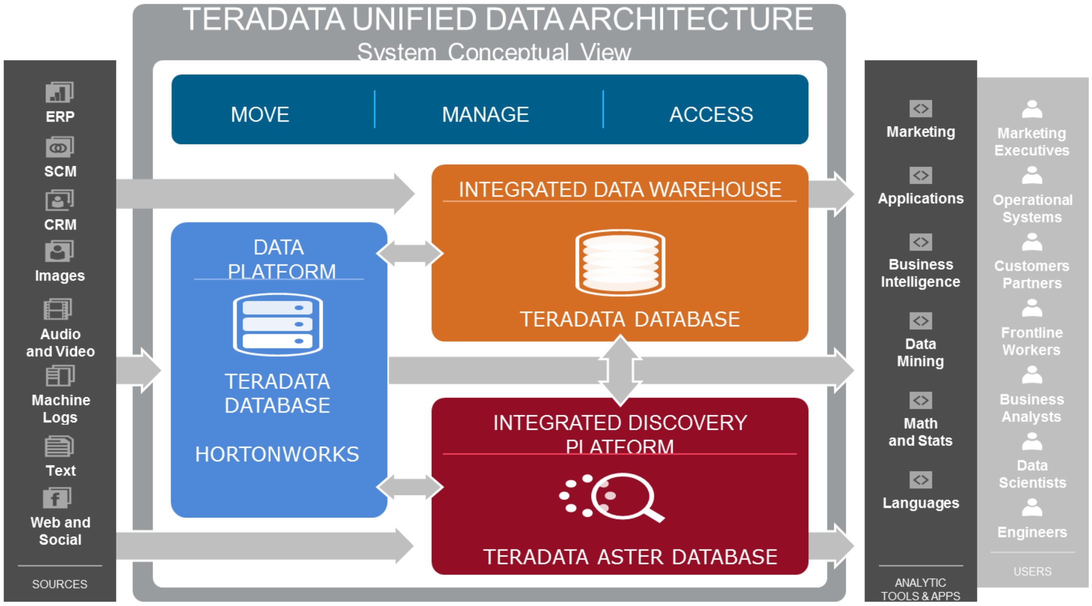

**Figure 4:** Teradata Unified Data Architecture

The Teradata Unified Data Architecture provides 3 distinct, purpose built data management platforms, each integrated with the others, intended with specialized needs:

- Integrated Data Warehouse - Teradata Database is the market-leading platform for delivering strategic and operational analytics throughout your organization, so users from across the company can access a single source of consistent, centralized, integrated data.
- Teradata Discovery Platform - Aster SQL-MapReduce delivers data discovery through iterative analytics against both structured and complex multi-structured data, to the broad majority of your business users. Pre-packaged analytics allow businesses to quickly start their data-driven discovery model, that can provide analytic lift to the SAS Analytics Platform.
- Data Capture and Staging Platform – Teradata uses Hortonworks Hadoop, an open source Hadoop solution to support highly flexible data capture and staging. With Hortonworks, Teradata has integrated it with a robust tools for system management, data access, and one-stop support for all Teradata products. Hadoop provides low-cost storage and pre-processing of large volumes of data, both structured and file-based.

With these three elements, the UDA provides flexibility to SAS users, enabling them to manage their data in a specific purpose platform, directed by the specific requirements of the analytic solution involved. This flexibility is critical for SAS users that are looking to integrate their analytic processing with enterprise data access services and their analytic directives into corporate decision making.

For most Teradata customers, they already utilize a Teradata EDW to gather data from multiple operational data sources. Integrated together, to provide a “single version of the truth” that can be interfaces from hundreds of downstream applications, or thousands of business users. However, this data is generally relational, organized into Tables, Rows and Columns, accessible by SQL. Depending on the data maturity that one may be working with, other platforms may be better suited:

- The Integrated Data Warehouse (IDW) is targeted at rigorous, highly structured, “ready for prime time” data that can be used by business and analytic users across the enterprise. IDWs address enterprise data readiness concerns, not only with the ability to scale to support more users, wider range of user applications, or larger volumes of data, but also scale to anticipate potential failures, keep running in the state of problems, and provide management and monitoring environment to ensure that the environment is continually to support data access from around the business. It is ideally suited for modeled and integrated data to support end-to-end optimization and operationalization

- The Discovery Platform is intended to for semi-structured data, still in need of analytic processing, but organized into functional areas sufficient that it can fit into a table oriented data structure. In this case, there's a tradeoff, where we are dropping some of the user concurrency requirements in exchange for high levels of analytic flexibility,
- The Hadoop Data Staging platform is intended to be an economical platform to capture and store varying types, the data that hasn't yet been fully structured, and doesn't yet need to be accessed for operational reporting across the enterprise. Frequently, it's used to collect sets of enormous files, like web logs, machine sensor outputs, or even web log outputs, all of which have some analytic value to them, but most importantly just need to be stored in a large clustered file system.

For some large companies, they utilize the different aspects of the UDA side by side, each with special purpose perspective on the end to end data lifecycle, evolving and maturing data from its "staged" landing area on Hadoop, right up to the "ready for action" integrated data warehouse environment.

## BRINGING IT ALL TOGETHER

Traditional systems lack the scalability and integrated architecture required to support big data analytics necessary in today's modern business environment. There has been a lack of harmony to combine data management and analytics, resulting in functional silos. These silos intensify the costs to maintain and manage, hurting the bottom line of the business. Our partnership enables users to:

- Effectively manage the ever-increasing data volumes, variety and velocity with the powerful Teradata platform coupled with SAS analytics and Hortonworks full Hadoop stack
- Minimize the cost of managing the disparate systems and technologies by offering an integrated architecture to help customers with their end-to-end data management and analytical process
- Deliver the BEST answer NOW to the right people, at the right time at any level within the organization

Teradata and SAS have joined forces to revolutionize your business by providing enterprise analytics in a harmonious data management platform to deliver critical strategic insights. By integrating SAS, Teradata and Hadoop, customers can take advantage of the best of breeds in the industry as illustrated in Figure 5.

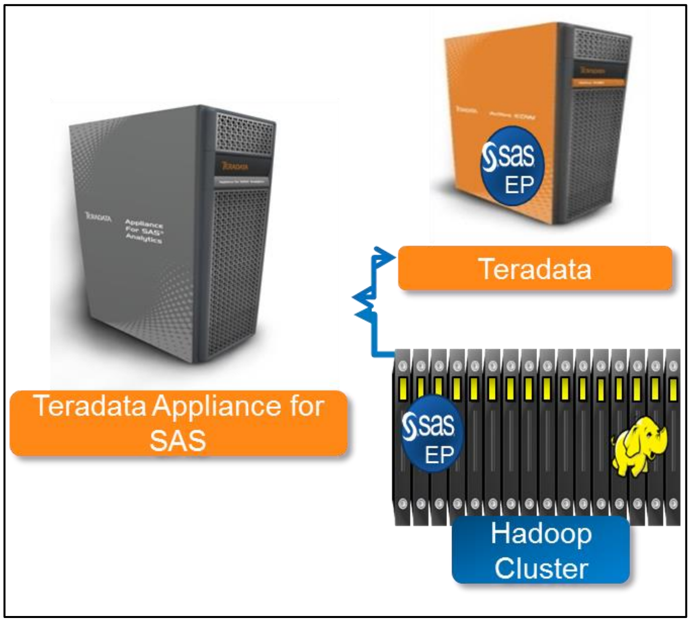

**Figure 5:** Integrating SAS, Teradata and Hadoop

- Hadoop
  - As a staging area, all types of data can be captured, stored and refine
  - Data from Hadoop can be extracted to Teradata Appliance for SAS - for data visualization (SAS Visual Analytics for Teradata) and for model development (SAS HPA Products as listed above)
- Teradata Platform
  - Leverage the integrated data warehouse to feed downstream analytical systems
  - Execute SAS in-database analytics such as data preparation (SAS® Data Quality Accelerator for Teradata, SAS® Code Accelerator for Teradata)
  - Develop data models (SAS® Analytics Accelerator for Teradata) and deploy the models (SASI® Scoring Accelerator for Teradata) in-database
- SAS Analytics
  - Improve performance by leveraging SAS EP multi-threading capability for parallel data movement
  - Combine the power of SAS in-database and in-memory processing for end-to-end analytic lifecycle

Customers can take advantage of this unique offering that combines the strengths of Teradata, SAS and Hadoop – the best of breeds in the industry. Organizations will be able to provide more insights from big data across the enterprise - enabling knowledge workers with faster analysis and executives to make more informed decisions.

## CONCLUSION

For big data analytics projects, one size **does not** fit all. Teradata and SAS have joined forces to revolutionize your business by providing enterprise analytics in a harmonious data management platform by applying advanced analytics “inside” the database or data warehouse where the vast volume of data is fed and resides. The key technologies that dramatically improve and increase performance when analyzing big data are “in-database” and “in-memory” analytics.

Depending on the nature of the data being analyzed, the ideal platform for storage may vary significantly.

- For big data that is file oriented, flat, or unstructured, ideal storage may be to utilize a large Hadoop cluster node distributing the data across several nodes in a cluster.
- For big data that is semi-structured, like huge flat files or loosely connected together, ideal platform may be to utilize a discovery platform analytic data mart that can be used to enrich and transform the data to derive higher value data elements.
- For big data that is highly structured, highly governed, used for operational decision making by hundreds across the organization, the ideal platform maybe an enterprise data warehouse, structured in a scalable integrated data warehouse.

By integrating SAS, Teradata and Hadoop, we **deliver** a cohesive environment that allows organizations to **manage ALL** your data, **simplify ALL** your data management and analytical processes and **empower** business executives to make real-time decisions for big data analytics.

## CONTACT INFORMATION

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

Name: John Cunningham
Enterprise: Teradata
E-mail: john.cunningham@teradata.com

Name: Tho Nguyen
Enterprise: Teradata
E-mail: tho.nguyen@teradata.com