PharmaSUG 2017 - Paper TT11

Tips and Best Practices using SAS® Analytics

Paul Segal, Teradata Corporation, San Diego, California

Tho Nguyen, Teradata Corporation, Raleigh, North Carolina

ABSTRACT

Many analytical computing solutions and large databases use new techniques because they provide significant performance improvements over the traditional methods. With SAS® in-database and in-memory analytics for Teradata, SAS users have the ability to prepare the data, 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. This paper discusses some tips and best practices with SAS/ACCESS, SAS formats, data quality, DS2, model development, model scoring, Hadoop and Visual Analytics - all integrated with the data warehouse

.

INTRODUCTION

Organizations are collecting more data than ever before, and it is presenting great opportunities and challenges to analyze ALL of that complex data in a timely manner. Trends in analytics and data management, along with heightened regulatory and governance burdens, demand new, innovative approach that can quickly transform massive volumes of data into strategic insight.

This paper will cover the following topics:

- SAS® Analytics for Teradata
  - In-database analytics
  - In-memory analytics
- Teradata Appliance for SAS®
- Hadoop in the data architecture

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.

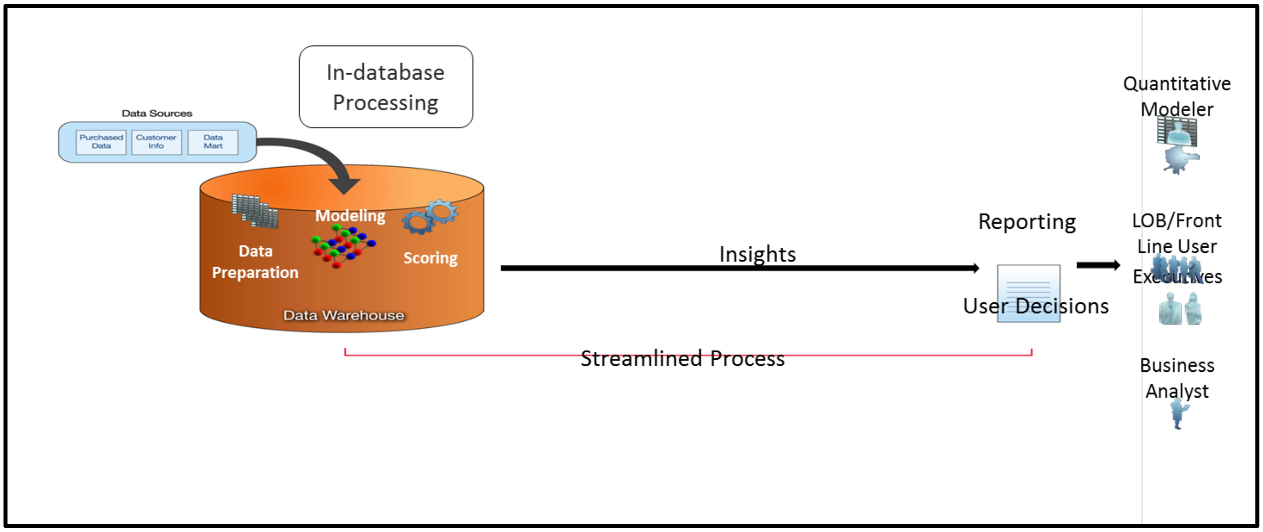

Figure 1: In-database processing: Minimize data movement and redundancy

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
    - Including PROC TRANSPOSE, allowing the transposition to occur in the Teradata database

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-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.

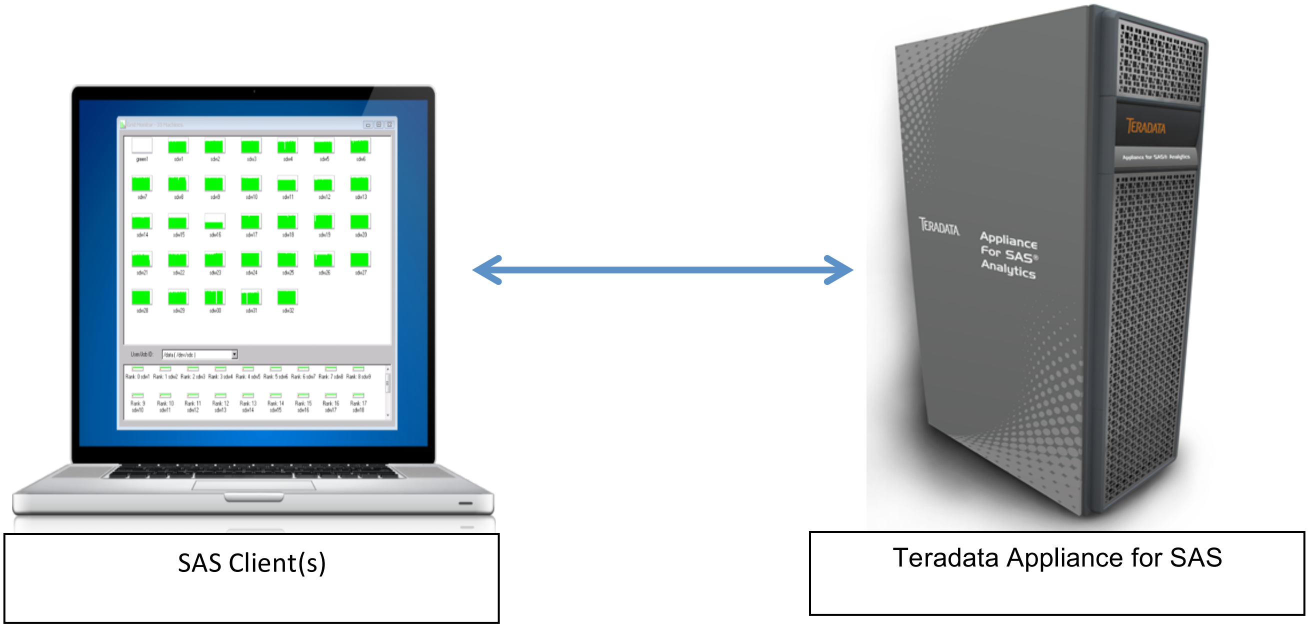

Figure 2: In-memory processing

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

The Teradata® Appliance for SAS is specifically for SAS High-Performance Analytics Products and SAS® Visual Analytics and Visual Statistics, 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 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

- **SAS Visual Analytics and Visual Statistics** - Explore massive volumes of data to quickly to visualize and uncover patterns and trends for further analysis
- **SAS 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 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, 128-768GB 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.

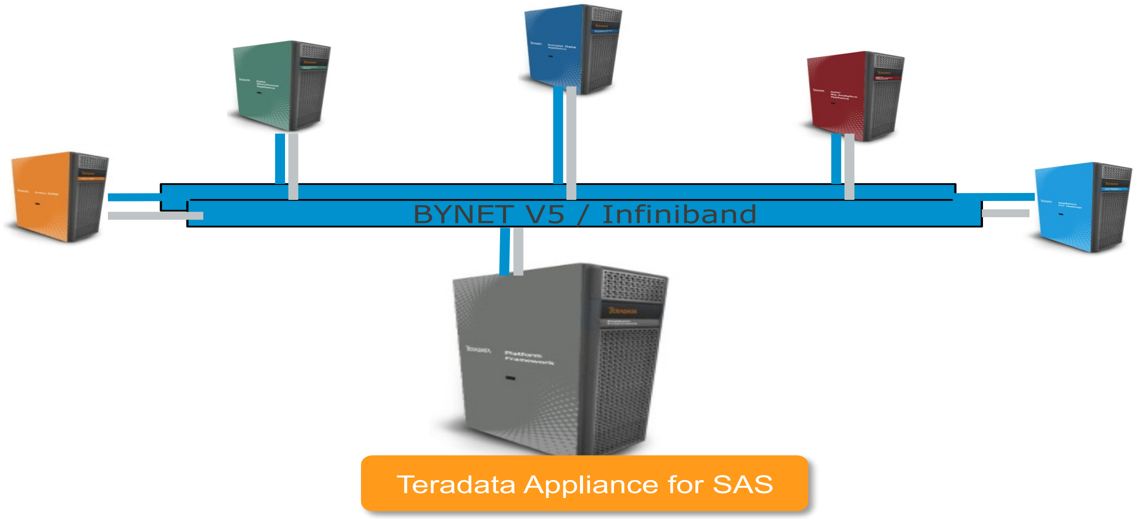

Figure 3: Teradata Platform Family Connects with Teradata Appliance for SAS

The Teradata Appliance for SAS 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 leverage 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 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 integrating Hadoop with the data warehouse.

CONCLUSION

In-database and in-memory processing are powerful and innovative approaches to managing large volumes of data without having to copy or move the data. By leveraging the power of the Teradata data warehouse and MPP architecture, many of the SAS complex computations can be executed in-database. In-database processing can

- Streamline analytics work flow
  - Minimize data preparation
  - Accelerate data discovery
- Increase performance
  - Reduce data movement
  - Leverage the MPP architecture for faster processing
- Improve data integrity
  - Enable data governance
  - Minimize information latency

CONTACT INFORMATION

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

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

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