SAS® In-Database Analytics for Teradata: Stop the Data Movement

Paul Segal, Teradata Corporation, San Diego, California

Tho Nguyen, Teradata Corporation, Raleigh, North Carolina

John Cunningham, Teradata Corporation, Danville, California

ABSTRACT

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 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 SAS in-database analytics for Teradata and some of the best practices adopted by our customers using Base SAS, SAS/STAT and SAS® Enterprise Miner.

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. In this volatile and competitive economy, there has never been a bigger need for proactive and agile strategies to overcome these challenges by applying the analytics directly to the data rather than shuffling data around. In addition, 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.

SAS and Teradata are addressing these challenges by moving analytic tasks closer to where the data reside with the integration of SAS analytics and the enterprise data warehouse (EDW) from Teradata. Performing these tasks inside Teradata will dramatically reduce the bottlenecks that result from moving data over a network. In addition, with the improved performance run times and reduction in redundant data, analytic models can be developed and deployed faster, turning the data into strategic insights.

This paper will cover the following topics:

- In-database analytics process
- Leveraging in-database analytics with Teradata
- Best practices and customer successes

IN-DATABASE ANALYTICS PROCESS

SAS and Teradata are the pioneers delivering in-database analytics solutions. Over the past 7 years, the adoption of in-database analytics has grown significantly and we have developed several programs to address these needs. 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 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.

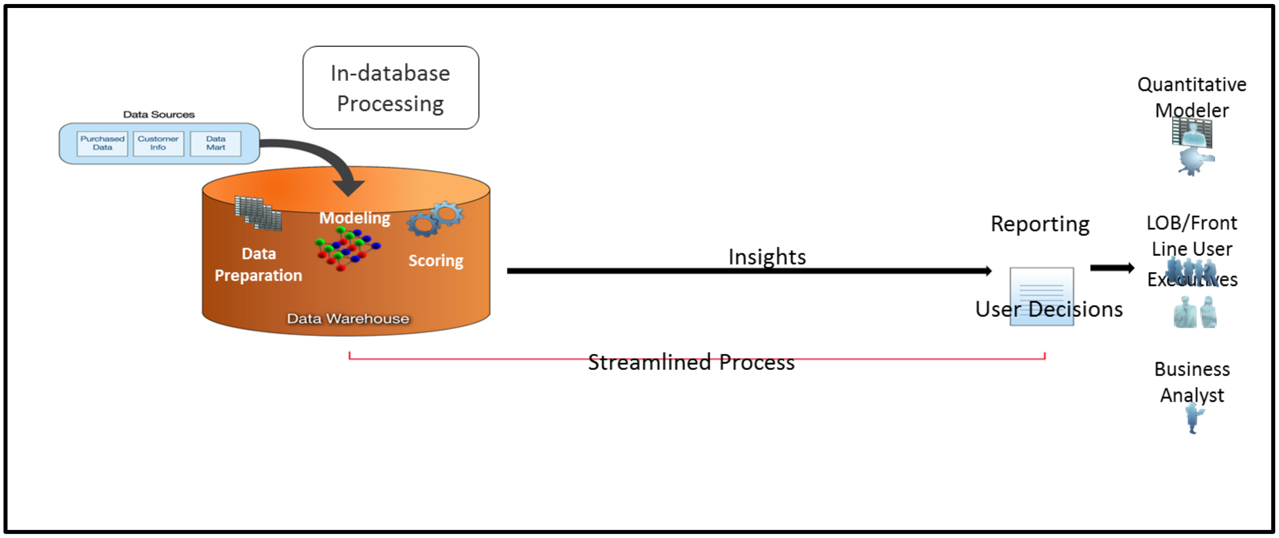

Figure 1: In-database processing: Minimize data movement and redundancy

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.

LEVERAGING SAS IN-DATABASE ANALYTICS WITH TERADATA

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 accelerates 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 known as advantage programs.

- **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.

BEST PRACTICES AND CUSTOMER SUCCESSES

Over the years, SAS and Teradata Center of Excellence have collected a number of best practices from various customer engagements and implementations. Below are just a few best practices that have worked well for our customers.

- Start small, then grow and expand – focus on proving the value by selecting a problem with long processing time and large volumes of data. For example, you may want to explore the power of PROC FREQ in-database and see the improved performance it can achieve. Then, move to other PROCS and products within the SAS portfolio.
- Consider a data lab – otherwise known as a ‘play pen’ or ‘sand box’ which is an area to explore and examine new ideas and possibilities by combining new data with existing data to create experimental designs and ad-hoc queries without interrupting the production environment. This is ideal to test and experiment the in-database functionality.
- Not everything can be executed in-database – combine the traditional method with the in-database approach since not all PROCS or tasks can be run the data warehouse. Consider executing as much in-database processing as possible and minimize the download of full tables into SAS.
- Apply options to monitor the communication between SAS and Teradata, and to encourage in-database execution.
- Utilize the Data Set Builder for SAS to prepare the data for analysis. This product builds the correct SQL syntax which can minimize the time to construct the data. It also allows code stream execution and ease of sharing projects.
- Exploit explicit Pass-through via PROC SQL as opposed to implicit PROC SQL. Best practice is to use explicit pass-through (EPT) particularly for non-simple queries. This approach reduces debugging time when in-database failures occur.

- Leverage the SAS Macro facility. The use of SAS macros is preserved in the SAS and Teradata integration. The SAS macro layer is interpreted first, which implies explicit pass-through can also utilize the facility. Finding ways to utilize the SAS Macro facility continues to be an efficiency standard as was true prior to the SAS and Teradata integration

The partnership has many customer successes and here are a few using in-database analytics.

- A larger retailer estimates elimination of 6 terabytes of redundant data.
- A major financial services organization plans to leverage the efficiencies of SAS In-Database processing with their Teradata system to increase SAS analytic processing speed ten-fold. This organization also estimates a reduction in model deployment time (from development to delivery) in terms of hours versus weeks. The reduced total cost of ownership (TCO) these efficiencies generate results in higher return on investment (ROI).
- A multinational Internet consumer-to-consumer corporation reduces processing time by a factor of 4.

Ultimately, the in-database approach has shown many benefits and process improvements.

CONCLUSION

In-database processing is a powerful and innovative approach 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